Suivi de l'activité d'une personne à partir de capteurs multi-modalités préservant l'anonymat dans un cadre de détection et prévention des chutes chez les personnes âgées

Imen Halima

THÈSE DE DOCTORAT DE

UNIVERSITÉ DE RENNES 1

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, image, vision*

Par
Imen HALIMA

Suivi de l'activité d'une personne à partir de capteurs multi-modalités préservant l'anonymat dans un cadre de détection et prévention des chutes chez les personnes âgées

Thèse présentée et soutenue à Rennes, le 03 mars 2021

Thèse N° :

Unité de recherche :

Laboratoire Traitement du Signal et de lImage (LTSI), UMR Inserm 1099 et Laboratoire Informatique et Télécom de l'ECAM

Rapporteurs avant soutenance :

Paul Checchin	Professeur à l'Université Blaise Pascal Clermont
François Charpillet	Directeur de recherche, Inria Grand-Est, France

Composition du Jury :

Président(e) :	Prénom NOM	Fonction et établissement d'exercice (à préciser après la soutenanc
Examinateurs :	Christine FERNANDEZ-MALOIGNE	Professeure à lUniversité de Poitiers
	Paul CHECCHIN	Professeur à l'Université Blaise Pascal Clermont
	François CHARPILLET	Directeur de recherche, Inria Grand-Est, France
	Jean-Marc LAFERTE	Enseignant-chercheur, ECAM Rennes
	Alain-Jèrôme FOUGERES	Enseignant-chercheur, HDR, ECAM Rennes
Dir. de thèse	Jean-Louis DILLENSEGER	Maitre conférence, LTSI-UMR 1099/ Inserm, Université de Rennes

Invité(s) :

Vincent GAUTHIER Directeur de Neotec Vision

TABLE DES MATIÈRES

TABLE DES FIGURES

Liste des tableaux

INTRODUCTION

Selon l'organisation mondiale de la santé (OMS), la chute est un problème majeur pour la santé publique, en particulier chez les personnes âgées. En effet, la chute est la deuxième cause de mortalité pour les seniors de plus de 65 ans. Selon le Baromètre santé, les chutes surviennent principalement à domicile (78% des chutes). Les blessures graves dues aux chutes peuvent entrainer la mort de la personne en cas d'absence de soins et de soutien immédiat. Par conséquent, un système de détection de chutes peut contribuer à améliorer la qualité de vie des personnes âgées indépendantes.

C'est dans ce contexte que notre thèse est venue renforcer scientifiquement les travaux menés dans le cadre d'un projet ANR (Agence Nationale de la Recherche) intitulé PRuDENCE. Ce projet est une collaboration entre l'entreprise NeoTec-Vision, spécialisée dans la vision par ordinateur, l'ECAM Rennes - Louis de Broglie, école d'ingénieurs généralistes, le laboratoire de traitement du signal et de l'image (LTSI, INSERM U1099) et l'Université de Technologies de Troyes UTT.
L'objectif de cette thèse est d'extraire des paramètres aidant à identifier l'activité de la personne à partir d'un dispositif à bas coût, afin de fournir des indices pour la détection de la chute en premier temps et la détection de la fragilité de la personne en second temps. Ce dispositif doit fonctionner d'une façon autonome, en temps réel et de jour comme de nuit en préservant l'anonymat. Pour ces raisons, nous avons choisi d'équiper ce système d'une paire de caméras : une caméra de profondeur associée à une caméra thermique de basse résolution.

Notre recherche apporte des contributions sur deux axes principaux : 1) un suivi temporel de la personne en incorporant les modalités de profondeur et thermique et 2) une estimation des paramètres qui serviront au suivi des activités à des fins de prévention de la fragilité de la personne.

Le premier axe de travail consiste à suivre la personne en se basant sur les images de profondeur et thermique. Plus particulièrement, nous nous sommes intéressés au suivi de la tête car sa vitesse verticale est importante en cas de chute. De plus, c'est la partie supérieure du corps, donc la zone la moins souvent occultée, elle est non déformable et sa température est souvent la plus élevée du corps. Pour réaliser le suivi de la tête, nous

commençons par une étape de soustraction de fond pour segmenter la silhouette de la personne. Par la suite l'algorithme de suivi est appliqué sur les images de premier plan qui contiennent l'ellipse englobante de la tête. Cet algorithme estime, à chaque image, la position de l'ellipse qui englobe la tête. Pour effectuer le suivi de la tête, nous nous sommes orientés vers le filtrage particulaire qui se base à chaque instant sur l'état précédent et les observations à l'instant présent. L'idée de ce filtrage est de créer plusieurs hypothèses de l'état actuel ("particules") associées à des poids. Chaque poids, calculé à partir de probabilités estimées sur les images, donne un niveau de confiance accordé à la particule. La position retenue peut être soit la particule de poids maximal, soit une moyenne pondérée (par les poids) des particules. Dans un premier temps le filtre a été utilisé séparément sur la modalité thermique puis celle de profondeur. Deux types d'images sont traités : image de profondeur et image thermique qui ont été capturées à partir d'un capteur Kinect v1 et Lepton Flir 1 respectivement. La fusion de ces images a amélioré les résultats de suivi.

Sur le deuxième axe de cette thèse, un système de reconnaissance de la posture de la personne est développé en se basant sur un apprentissage profond. L'idée est d'estimer les paramètres qui serviront au suivi des activités à des fins de prévention de la chute en appliquant le Deep learning sur des séquences d'images thermiques et de profondeur. Ce projet a été initié lors d'une mobilité de 3 mois au LIST, Université du Sud-Est, Nankin, Chine. Le système localise et détecte les postures de la personne en utilisant le SSD (Single Shot Detector). Cette méthode est supervisée, basée sur le réseau convolutif VGG16. Dans notre contexte, nous avons choisi quatre postures à détecter (debout, assis, allongé sur un lit/canapé et allongé sur le sol). Pour avoir une classification performante, nous avons créé une base de données à l'aide de multiples acquisitions de séquences sur plusieurs lieux (un living lab à l'UTT et un appartement à l'ECAM de Rennes) et avec sept acteurs différents.

Ce manuscrit est organisé en trois parties principales : 1) Une étude bibliographique sur les systèmes de détection de chutes et une définition du contexte général de notre projet, 2) un développement d'une méthode de suivi de la personne avec des capteurs basés sur la vision et 3) une présentation d'un modèle d'analyse de la posture de la personne à des fins de prévention de la fragilité des personnes âgées.
La première partie contient deux chapitres. Le premier chapitre présente les causes et les conséquences de la chute. Ensuite nous proposons une étude détaillée des systèmes de détection de chute et des systèmes de prévention de la fragilité de la personne existants. Dans le deuxième chapitre, nous commençons par présenter le contexte général de cette

thèse ainsi que le dispositif utilisé, puis nous expliquons la phase de création des données à traiter par la suite.
La deuxième partie est organisée en trois chapitres. Le premier chapitre consiste à définir le principe des méthodes de suivi, plus particulièrement, le principe de filtrage particulaire. Dans le deuxième chapitre, nous détaillons une méthode de suivi de la tête, basée sur le filtrage particulaire et dont les résultats sont appliqués sur des images de profondeur. Enfin nous montrons, dans le troisième chapitre, l'intérêt de la fusion des informations de profondeur et thermiques en comparant les résultats issus de la fusion avec ceux obtenus à partir de l'information de profondeur seule.
La troisième partie se compose de deux chapitres. Dans le premier, nous expliquons le principe des méthodes d'apprentissage profond, en détaillons les techniques utilisées pour détecter et localiser les postures de la personne. Le deuxième chapitre présente notre système de classification de postures et son application sur les images de profondeur et thermiques. Nous présentons, par la suite, l'intérêt de la fusion pour améliorer les résultats obtenus.
Nous achevons ce manuscrit par une conclusion qui résume les travaux présentés dans ce mémoire ainsi que les perspectives envisagées à court et long terme pour améliorer les performances du système développé dans le cadre du projet PRuDENCE.

MÉTRIQUES D'ÉVALUATION

L'objectif de ce chapitre introductif est de définir les métriques d'évaluation utilisées dans cette étude et évoquées tout au long de ce rapport. Une métrique est un moyen de mesure quantifiable de la performance d'un modèle de classification et de détection.

D'abord, nous commençons par présenter les différentes métriques du modèle de classification binaire. Puis, nous appliquons ces métriques sur les modèles n-aire. Enfin, nous détaillons les métriques dédiées à la détection d'objets.

A Classification d'objets : Modèle binaire

Un modèle binaire contient exactement deux classes : une classe positive (par exemple chute) et une classe négative (non-chute). Les métriques de ce modèle sont basées sur ces notions basiques :

- Vérité terrain : elle contient l'ensemble des événements réels étiquetés à la main (étiquette positive et étiquette négative).
- Vrai Positif (TP : True Positive) : lorsque l'étiquette et la prédiction sont positives.
- Vrai Négatif (TN : True Negative) : lorsque l'étiquette et la prédiction sont négatives.
- Faux Positif (FP : False Positive) : lorsque l'étiquette est négative alors que la prédiction est positive.
- Faux Négatif (FN : False Negative) : lorsque l'étiquette est positive alors que la prédiction est négative.

A.1 Matrice de confusion

La matrice de confusion, aussi appelée matrice d'erreur, est un moyen pour mesurer les performances d'un modèle de classification autre que les métriques d'évaluation. C'est une visualisation tabulaire des prédictions du modèle par rapport à la vérité terrain et elle peut s'étendre à un nombre de classes supérieures à deux. On note que chaque ligne et chaque colonne de la matrice de confusion représentent respectivement les classes réelles

et les classes prédites.

		Prédiction	
		Positif	Négatif
Réel	Positif	TP	FN
	Négatif	FP	TN

FIGURE 1 – Matrice de confusion.

Comme l'illustre la Figure 1, les cellules en diagonale de cette matrice indiquent la prédiction correcte des différentes classes, tandis que les autres cellules indiquent les erreurs de prédiction (FP et FN). Elle est appelée "matrice de confusion" car elle permet de repérer facilement les confusions entre les deux classes.

A.2 Précision

La précision, également nommée la valeur prédictive positive, est le rapport entre le nombre de TP et la somme des FP et TP (Équation 1).

$$Précision = \frac{TP}{TP + FP} \tag{1}$$

Il faut souligner que la valeur numérique de cette notion est comprise entre 0 (aucune vraie alarme) et 1 (aucune fausse alarme). Cette métrique représente le nombre des prédictions correctement prédits comme positives sur le total des éléments positifs (Figure 2).

Prenons le cas des détections des chutes, la précision est le rapport des vraies détections de chute par rapport à l'ensemble des détections de chutes. Elle permet, ainsi, de fournir la probabilité des vraies alarmes. Par exemple, une précision de 0.5 signifie que la moitié des alarmes correspond à des chutes réelles et l'autre moitié correspond aux fausses alarmes (1 faux positif pour chaque chute détectée).

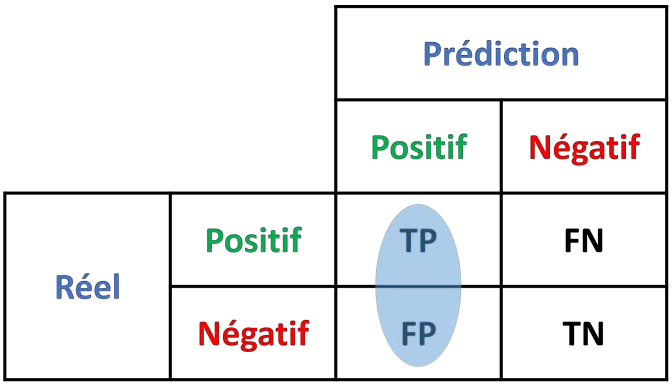


FIGURE 2 – Précision.

A.3 Spécificité (Specificity)

La spécificité, aussi appelée le taux négatif réel, est la proportion d'événements autres que la chute qui sont correctement détectés (Équation 2).

$$Spécificité = \frac{TN}{TN + FP} \tag{2}$$

Il s'agit du nombre d'éléments correctement identifiés comme négatifs sur le total des négatifs (Figure 3).

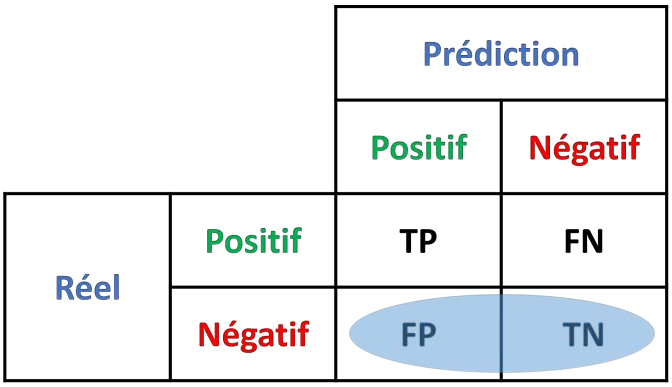


FIGURE 3 – Spécificité.

La spécificité quantifie la probabilité d'éviter un faux positif (fausse alarme). Le contraire de la spécificité est le taux de faux positifs, qui est la proportion d'événements

autre que la chute détectée par erreur comme des chutes (Équation 3).

$$Taux\ de\ faux\ positifs = \frac{FP}{TN + FP} = 1 - Spécificité \tag{3}$$

A.4 Rappel (Recall/sensitivity)

Le rappel, connu aussi sous le nom de sensibilité ou bien taux positif réel, est le rapport entre le nombre de TP et la somme des TP et FN (Équation 4).

$$Rappel = \frac{TP}{TP + FN} \tag{4}$$

Il définit le nombre des prédictions correctement identifiées comme positives sur le total des événements réellement positifs de la base de données. Il représente la partie supérieure de la matrice de confusion (Figure 4).

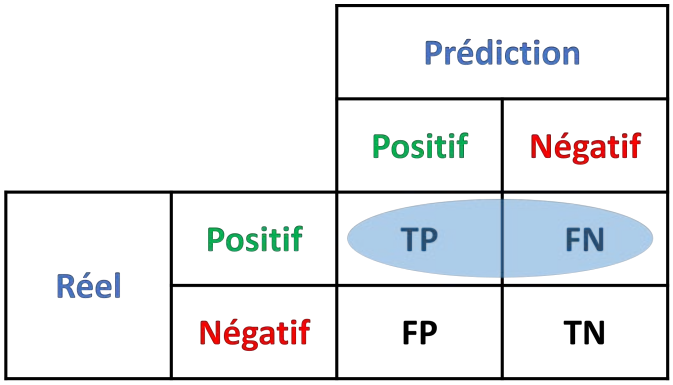


FIGURE 4 – Rappel.

Le rappel s'avère le critère le plus utilisé dans la littérature. Il définit la proportion de chutes qui sont correctement détectées. Le contraire de ce critère (i.e. le contraire de la sensibilité) est relatif au taux d'échec (taux de faux négatifs) qui quantifie la proportion de chutes non détectées (Équation 5).

$$Taux\ de\ faux\ négatifs = \frac{FN}{TP + FN} = 1 - Rappel \tag{5}$$

A.5 Score F1

Le rappel donne des informations sur la performance d'un modèle en ce qui concerne les faux négatifs (combien de détections ont été ratées). Tandis que la précision donne des informations sur la performance du modèle en ce qui concerne les faux positifs (combien de prédictions ont été correctement détectées). Le modèle est proche de l'idéal s'il a un minimum de nombres de faux négatifs (rappel proche de 100%) et faux positifs (précision proche de 100%). Pour évaluer le déséquilibre dans le cas de base déséquilibrée, le score F1, connu aussi par le coefficient de Dice, a été introduit pour prendre en compte la précision et du rappel en utilisant la moyenne harmonique.

Considérons le cas d'une base de données déséquilibrée avec plus d'importance pour une seule classe, le choix du modèle le plus performant est basé sur le score F1 le plus élevé pour cette classe. Par exemple pour la détection des fraudes, il est plus important de prédire correctement une transaction comme frauduleuse que prédire la non frauduleuse.

Moyenne harmonique : C'est l'inverse de la moyenne arithmétique des inverses. Autrement dit, la moyenne harmonique sert à réduire l'impact des grandes valeurs sur les petites valeurs (Équation 6). Elle est comprise entre 0 et 1 :

$$F1 = \frac{2}{\frac{1}{Précision} + \frac{1}{Rappel}} = 2 * \frac{Précision * Rappel}{Précision + Rappel} = \frac{2\ TP}{2\ TP + FN + FP} \tag{6}$$

Selon ces termes, l'utilisation de la moyenne harmonique est plus intuitive pour compenser le déséquilibre harmonique des données. Prenons à titre d'exemple le cas d'un système de reconnaissance d'empreinte digitale qui se caractérise par une précision et un taux de rappel égaux respectivement à 1 et 0.2. Théoriquement, la performance devrait être très faible car le système ne reconnait que 20% des empreintes enregistrées, ce qui signifie qu'il est presque inutile. Sachant que la moyenne arithmétique de 1 et 0,2 est égale à 0,6 alors que la moyenne harmonique de ces deux valeurs est égale à 0.33. Pour cette raison le choix de la moyenne harmonique est plus raisonnable.

A.6 Justesse (Accuracy)

La justesse est la probabilité des prédictions bien classées. Elle évalue les éléments en diagonale de la matrice de confusion qui désignent les prédictions correctes (Figure 5).

La justesse est le rapport de la somme de TP et TN et le nombre total des cas examinés

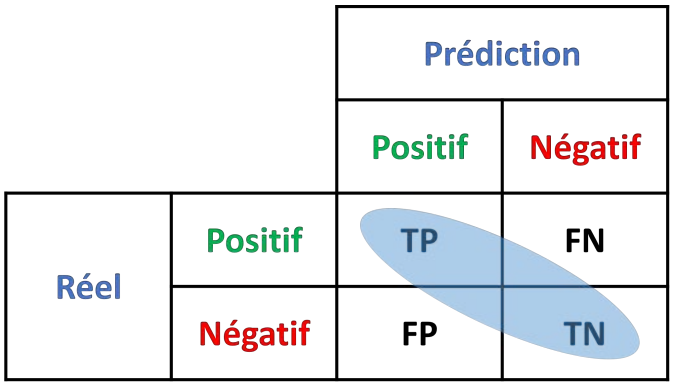


FIGURE 5 – Justesse.

(Équation 7) :

$$Justesse = \frac{TP + TN}{TN + FP + TP + FN} \tag{7}$$

Sa valeur est comprise entre 0 et 1, sachant qu'une valeur élevée indique que le modèle est meilleur.

B Classification d'objets : Modèle n-aire

Dans cette section, nous appliquons les métriques définies précédemment sur les modèles de classification à plusieurs classes, appelés "Classification n-aire". Elle présente un défi différent de celui de la classification binaire. En effet, elle comporte plus de deux classes. Nous prenons par exemple le cas d'une base de données contenant 3 classes différentes de fruits (Pomme, Orange et Poire). Chaque image de cette base est affectée à une et une seule étiquette.

B.1 Matrice de confusion

La matrice de confusion est une bonne technique pour évaluer les performances d'un algorithme de classification n-aire. En prenant l'exemple de trois classes de fruits, la Figurefig :matnaire illustre une matrice 3×3 :

Dans ce cas, les cellules en diagonale représentent le nombre des détections correctes de chaque classe, sachant que les autres cellules correspondent aux nombres de confusions

		Prédiction		
		Orange	Pomme	Poire
Réel	Orange	4	1	1
	Pomme	6	2	2
	Poire	3	0	6

FIGURE 6 – Matrice de confusion.

de la classe en colonne avec celle en ligne.

B.2 Précision

Pour la classifications n-aire, la précision est calculée pour chaque classe. Elle est traduite par le rapport entre le nombre de détections correctement prédites et l'ensemble des prédictions de cette classe.

En prenant l'exemple de la classe Orange (Figure 6), cette métrique est égale à 0.31 en divisant le nombre de vrais prédictions "4" par la somme des prédictions Orange "13" $(4 + 6 + 3)$. Nous déduisons que la classe Orange est correctement prédite dans les 31% des cas, et ainsi de suite pour les deux autres classes (Table 1).

Classe	Précision	Rappel	Score F1
Orange	0,308	0,667	0,421
Pomme	0,667	0,200	0,308
Poire	0,667	0,667	0,667

TABLE 1 – Précision, rappel et score F1.

B.3 Rappel

De même que la précision, chaque classe a un taux de rappel spécifique. Il est calculé en divisant le nombre de vraies prédictions par le nombre de cas réels. D'après l'exemple de la classe Orange, cette métrique est égale à 0.67. Cela signifie que ce classifieur a bien détecté 2/3 des Oranges comme Orange (Table 1).

B.4 Score F1

Comme indiqué dans la section A.4, le score F1 de chaque classe est la moyenne harmonique de la précision et du rappel de cette classe (Table 1).

C Détection d'objets

Après avoir détaillé les métriques d'évaluation des méthodes de classification d'objets, nous abordons dans cette section les métriques les plus utilisées pour la "Détection d'objets". Cette dernière localise plus précisément l'emplacement de cet objet dans l'image par une boîte englobante.

C.1 IoU (Intersection over Union)

IoU, également appelé indice de Jaccard, est une mesure qui quantifie la similarité entre la boîte englobante détectée et la boîte "étiquette"(celle de la vérité terrain) (Figure 7.10). Cet indice est égal au rapport entre l'intersection et l'union de ces deux boîtes (Équation 8) :

$$IoU = \frac{intersection}{union} = \frac{A \cap B}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{8}$$

Un seuil d'IoU est fixé pour chaque base de données pour différencier le vrai positif de faux positif. Par exemple, un seuil de $0,5$ signifie que la boîte englobante est un vrai positif lorsque son intersection avec la vérité terrain dépasse la moitié de leur union. Ce critère est utilisé pour calculer le nombre d'objets correctement détectés et le nombre de faux positifs qui ont été générés. Le score d'IoU varie entre 0 et 1. Plus les deux boîtes sont proches, plus ce score est élevé (Figure 8).

C.2 Courbe de rappel-précision

La courbe de rappel-précision (Precision Recall Curve PRC), illustrée dans la Figure 9, représente la performance du modèle. L'abscisse et l'ordonnée désignent respectivement le rappel et la précision. Chaque point de la courbe est tracé en fonction des scores de

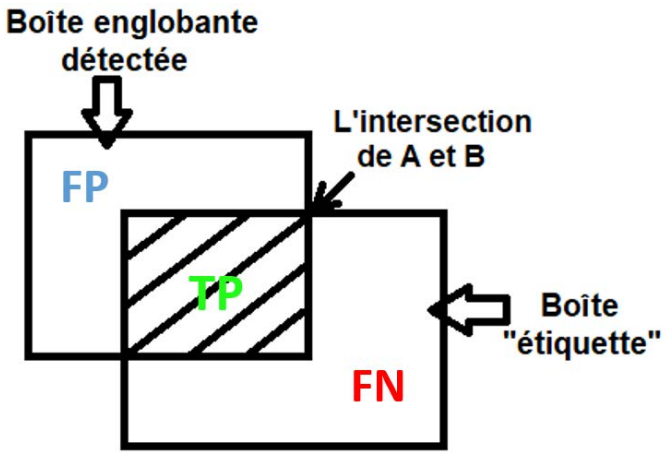


FIGURE 7 – Définition d'IoU.

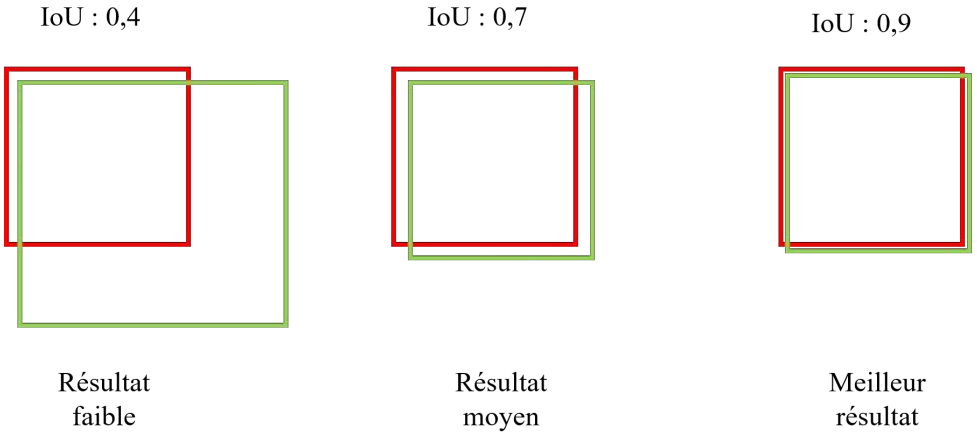


FIGURE 8 – Exemple de différents cas d'IoU.

prédiction (un seuil d'IoU est déjà fixé pour différencier un vrai positif d'un faux positif). Cette courbe est strictement décroissante. En effet le rappel augmente quand la précision diminue. Un modèle est plus efficace lorsque sa courbe (Figure 10 : courbe A) dépasse les autres (située en haut à droite).

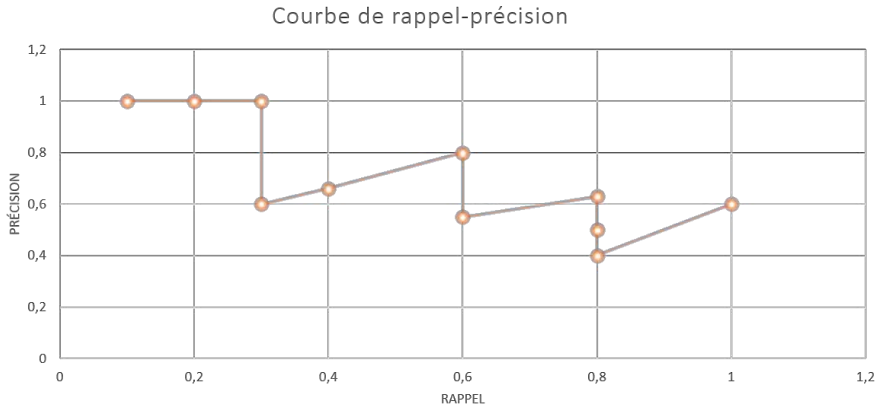


FIGURE 9 – Courbe de rappel-précision.

C.3 Précision moyenne

La courbe PRC évalue la performance de plusieurs modèles. Néanmoins, elle peut avoir des formes particulières (sous forme de dents de scie), ce qui rend difficile la comparaison de courbes croisées. Dans ce contexte, une métrique numérique, appelée "Précision Moyenne (Average Mean AM), est introduite afin d'ajuster cette courbe. Cette métrique interpole la précision à chaque intervalle de rappel. Une des premières solutions adoptées pour ajuster la courbe est l'interpolation à 11 points. Ces points sont fixés en divisant l'intervalle [0,1] par un pas de 0.1. A chaque sous intervalle de rappel, la valeur de précision interpolée $p_{interp}(r)$ est la valeur de précision maximale à droite de cet intervalle (Figure 11).

Par conséquent, les variations de la forme particulière de cette courbe (forme de dents de scie) sont réduites. La définition générale de la précision interpolée sur 11 points consiste à moyenner la somme des précisions interpolées (Équation 9) :

$$AP = \frac{1}{11} \sum_{r \in [0,0.1,0.2,\ldots,1]} p_{interp}(r) \tag{9}$$

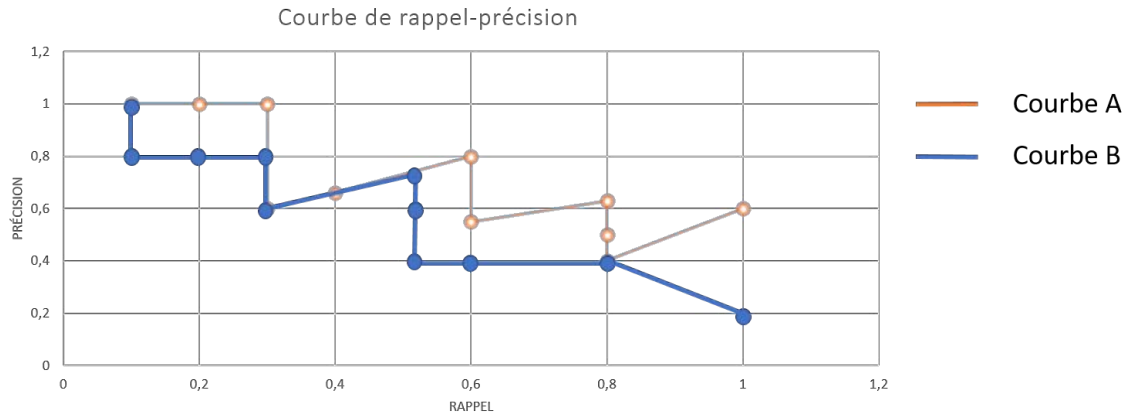


FIGURE 10 – Comparaison de deux courbes de rappel-précision.

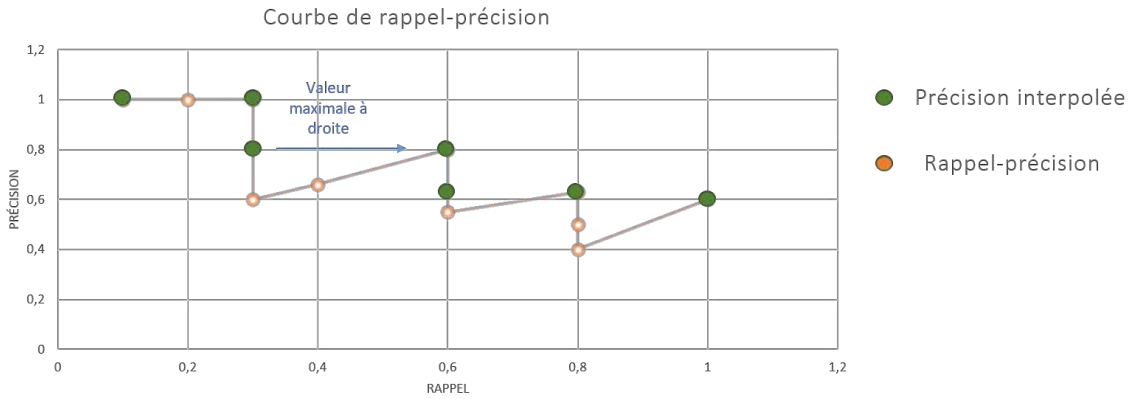


FIGURE 11 – Précision moyenne interpolée sur 11 points.

$$\text{avec } p_{interp}(r) = \max_{\tilde{r}>r} p(\tilde{r})$$

Une autre technique plus élaborée consiste à ajuster cette méthode d'interpolation en tenant compte notamment de la totalité de l'aire sous la courbe. En effet, l'AUC est calculée à l'aide des points interpolés précédemment (Figure 11).

Dans la Figure 12, la nouvelle courbe interpolée illustre le calcul de la précision moyenne qui est égale à la somme des aires de différents niveaux de précision.

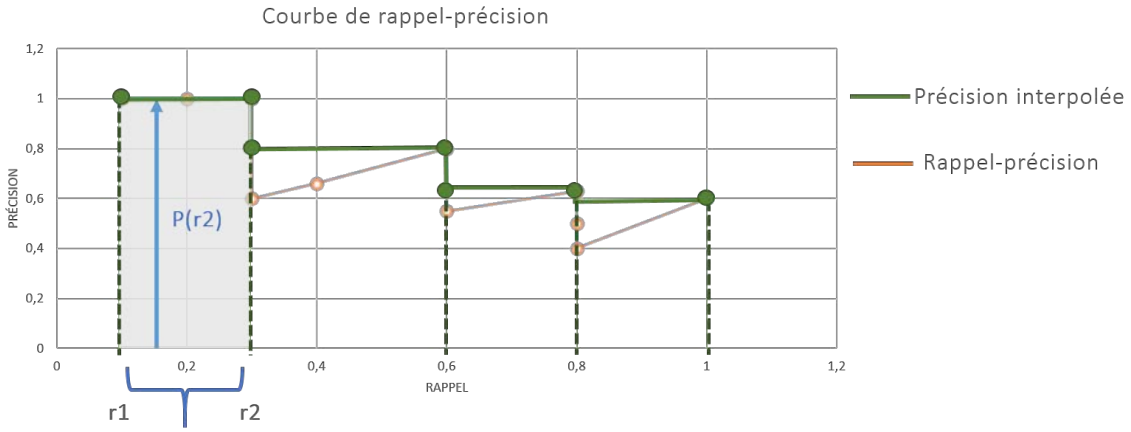


FIGURE 12 – Calcul de la précision moyenne.

L'Équation 10 illustre le calcul mathématique de cette aire :

$$AP = \frac{1}{11} \sum (r_n - r_{n-1}) p_{interp}(r) \tag{10}$$

C.4 Score mAP

Chaque image x d'une base de donnée peut contenir un ou plusieurs objets qui appartiennent à une ou plusieurs classes. La métrique de précision moyenne évalue la performance de chaque classe. En se basant sur la PA, la métrique mAP (mean Average Precision) est utilisée pour examiner l'état général du modèle. Elle est égale à la moyenne des PA de toutes les classes (C : nombre de classes) (Équation 11).

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{11}$$

Première partie

État de l'art et contexte général

Chapitre 1

État de l'art

Introduction

Selon l'organisation mondiale de la santé (OMS), la proportion de la population mondiale de plus de 60 ans doublera d'ici 2050 pour passer d'environ 11% à 22%. Le nombre absolu de personnes âgées de 60 ans et plus devrait augmenter pour passer de 605 millions à deux milliards au cours de cette période [1]. D'après l'Institut National de la Statistique et des Études Économiques (INSEE), le nombre des personnes âgées de plus de 60 ans en France passera de 24.8 millions à 31.9 millions entre 2015 et 2050 comme l'indique la Table 1.1

Année	**2015**		**2050**	
Tranche d'âge	De 60 à 74 ans	plus de 75 ans	De 60 à 74 ans	plus de 75 ans
Estimation du nombre de personnes âgées en France	15,5 millions	9,3 millions	15,9 millions	16 millions

Table 1.1 – Estimation du nombre des personnes âgées [2].

Dans les pays occidantaux, le nombre de personnes âgées qui ont perdu leur autonomie devraient être multiplié par quatre d'ici à 2050. De nombreuses personnes très âgées ne peuvent plus vivre seules car elles ont du mal à se déplacer, elles sont fragiles ou ont d'autres problèmes de santé physique ou mentale. Un grand nombre d'entre elles ont besoin d'une prise en charge (soins à domicile, soins prodigués au sein de la communauté, assistance pour les tâches de la vie quotidienne, logement en établissement spécialisé ou hospitalisation prolongée) [1].

Tenant compte de cette situation, il devient nécessaire d'adapter les domiciles pour

qu'elles vivent avec le maximum d'autonomie possible. Par contre, une des plus grandes craintes de personnes âgées qui se trouvent en autonomie est la chute, surtout si elle sont seules. En effet, en maison de retraite, une moyenne de 2-3 chutes est constatée par personne et par ans avec souvent des conséquences assez graves. Prenons l'exemple de la France, 9300 personnes de plus de 60 ans décèdent chaque année à cause des chutes. Celles-ci surviennent principalement au domicile (78% des chutes) et pendant la nuit (60% de ces chutes). Mais sans conséquences funestes, les chutes engendrent des conséquences physiques et psychologiques. Ainsi, il est primordial de trouver des solutions pour minimiser les risques et les conséquences dûs aux chutes des personnes âgées et construire des systèmes qui facilitent leur quotidien chez eux ou même dans les maisons de retraite.

Dans ce chapitre, nous analysons dans un premier temps les causes et les conséquences des chutes et par la suite nous expliquons les systèmes de détection et de prévention dédiés à la chute. Puis, nous détaillons le contexte dans lequel se situe cette thèse. Enfin, nous présentons les outils utilisés ainsi que les contributions.

1. 1 Facteurs de risque de la chute

Selon l'OMS, la chute est la deuxième cause des décès accidentels ou des décès par traumatisme involontaire dans le monde. C'est un problème de santé publique majeur, en particulier chez les personnes âgées [1].

Une étude statistique montre qu'un adulte sur trois âgé de 65 ans ou plus tombe au moins une fois par an. Par ailleurs, le risque de chute est multiplié par 20 après une première chute et le risque de décès s'accroit d'un facteur de quatre dans l'année qui suit la chute. L'impact et les conséquences d'une chute peuvent être réduits lorsque l'événement est détecté en temps réel et que des soins médicaux sont fournis rapidement [3].

Pour améliorer la qualité de vie des personnes âgées indépendantes, il faut d'abord identifier les conséquences des chutes, ensuite examiner les facteurs de risques et enfin proposer des solutions afin de réduire indirectement les dommages faits à la personne.

Les chutes peuvent avoir des conséquences physiques, psychologiques, sociales, financières et médicales néfastes, ainsi que des conséquences pour l'État et la communauté. Après une première chute, l'indépendance de la personne diminue à cause de la peur de tomber à nouveau, ce qui l'empêche de rester active. Par conséquent, ces personnes âgées ont besoin de soins permanents, ce qui coûte cher pour la société ainsi que pour l'État. Par

exemple, le centre de contrôle et de prévention des maladies (Centers for Disease Control and prevention CDC) a indiqué que le coût moyen d'une chirurgie due à une chute est de plus de 30000 dollars. La Figure 1.1 détaille les conséquences principales liées à la chute qui sont coûteuses à la fois sur le plan moral et sur le plan économique.

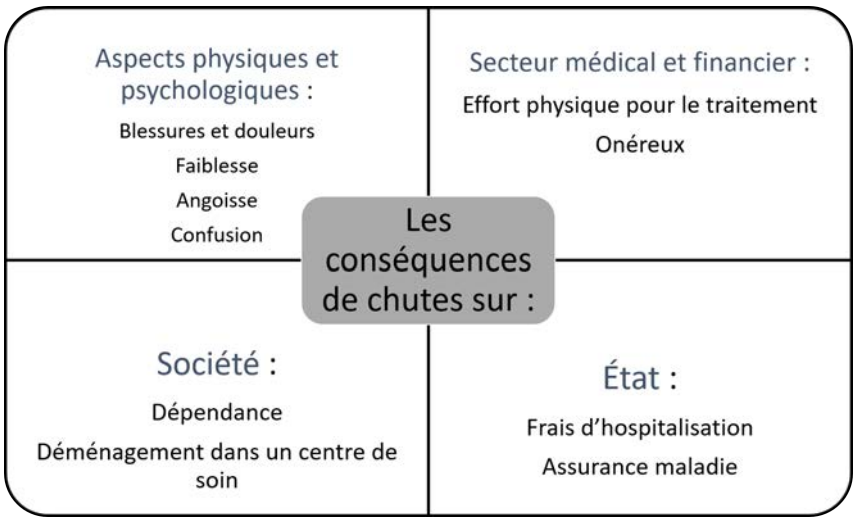


FIGURE 1.1 – Les préjudices de la chute sur plusieurs secteurs.

Plusieurs études ont montré que le risque de chute augmente considérablement à mesure que le nombre de facteurs de risque augmente. Dans ce contexte, la Figure 1.2 énumère les principaux facteurs de risque qui perturbent l'autonomie des personnes âgées. Ils sont classés en deux catégories, intrinsèques et extrinsèques. Dans ce document, nous avons identifié l'âge, le manque d'équilibre et la faiblesse musculaire comme facteurs intrinsèques et les marches d'escalier, le sol glissant et le manque d'équipement comme facteurs extrinsèques.

Selon ces facteurs, différents types de chutes peuvent être détectés. Ils sont classés en trois catégories : une chute molle en avant/arrière, une chute latérale (à droite/gauche) et une chute lourde (en cas de perte de verticalité rapide) comme indiqué dans la Table 1.2.

Lorsqu'une chute se produit, l'accélération du corps change soudainement. Pour mieux comprendre le mécanisme d'une chute, le changement d'accélération du corps est illustré dans la Figure 1.3 sous forme de trois phases. La première phase, qui se produit juste avant la chute, dure quelques millisecondes. Elle s'appelle le délai d'exécution. La deuxième phase est le moment de la collision, lorsque l'accélération du corps augmente rapidement.

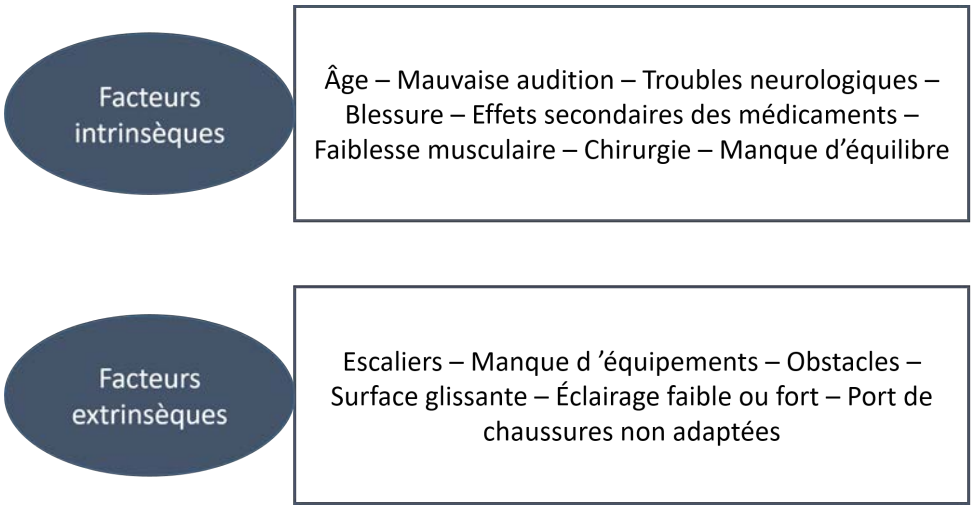


FIGURE 1.2 – Les facteurs de chutes.

Type de chute	**Facteurs**
En avant / en arrière	En présence d'obstacles / sol glissant / en montant les escaliers / en s'assoyant / en se mettant debout / en se penchant pour enlever quelque chose / mauvaise audition / manque d'équilibre
Latérale (à gauche / à droite)	En présence d'obstacle / sol glissant / en dormant sur un lit / en s'assoyant sur une chaise / mauvaise audition / manque d'équilibre
Lourde	Troubles neurologiques / manque d'équilibre / en présence d'obstacles

TABLE 1.2 – Types de chutes.

La personne tombe par terre et au moment de cette phase, des blessures graves peuvent se produire. La dernière phase (Post-chute) se déclenche juste après la chute. Pendant cette phase, l'accélération du corps diminue et la personne est en état de choc [4].

De nombreux travaux de recherche ont été menés pour réduire les conséquences des chutes et améliorer la qualité de vie des personnes âgées. Les chercheurs ont proposé un dispositif communicant qui s'intègre dans la vie de la personne dans l'objectif d'alerter en cas de chute. À chaque événement anormal, une alarme, contenant l'état de la personne, est envoyée aux soignants. Cette alarme est générée à l'aide d'un système de détection de chute automatique. Dans le cas d'un évènement critique, une décision sera envoyée aux personnes concernées (infirmier, proches) sous forme d'une alarme. La Figure 1.4 illustre

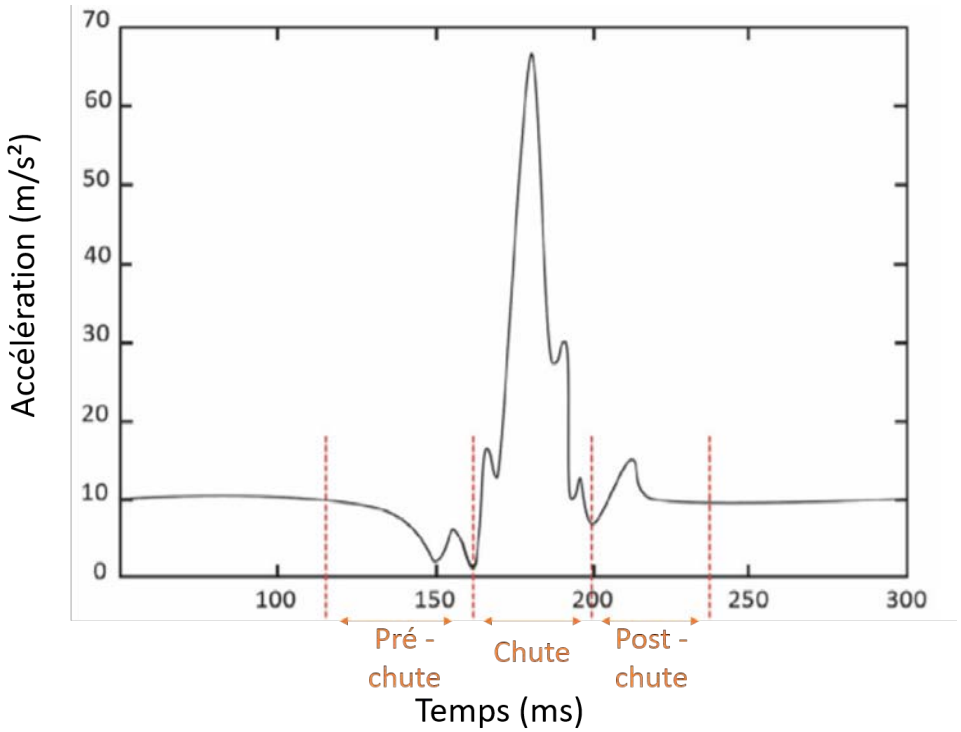


FIGURE 1.3 – Analyse temporelle d'une chute [4].

les différents éléments d'un détecteur de chute.

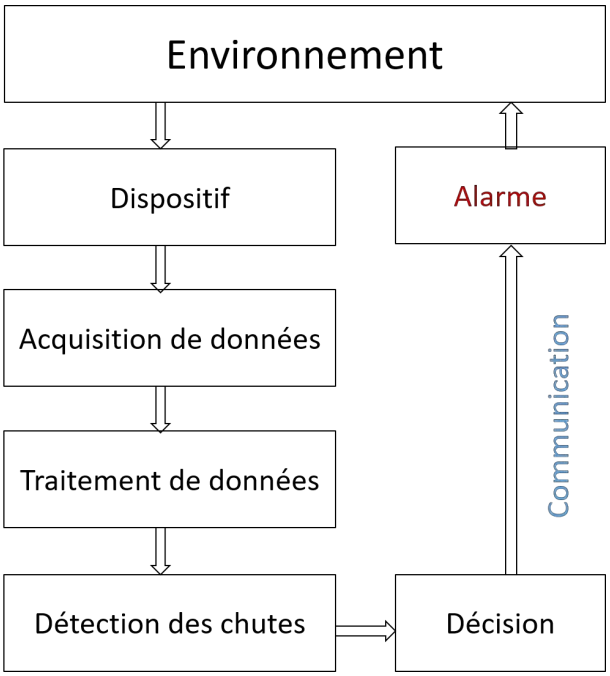


FIGURE 1.4 – Cadre général d'un système de détection des chutes.

1. 2 Systèmes de détection des chutes des personnes âgées

L'intégration d'un système de prévention et de détection de chutes (Fall Prevention FP, Fall Detection FD) est nécessaire pour envoyer des alarmes en cas d'urgence, diminuer le temps d'intervention et augmenter les chances de survivre.

Certains chercheurs ont classé les systèmes exsistants en utilisant deux variétés de critères de comparaison : types de chutes et types de capteur utilisé. Mubashir et al. [5] ont introduit différents types de chutes selon différentes positions, notamment en marchant, debout, couché dans un lit et assis sur une chaise. En outre, ils ont classé les approches FD en trois catégories en fonction de l'appareil utilisé : les appareils portables, les capteurs d'ambiance et les caméras. Cependant, Igual et al. [6] ont identifié trois autres types de chute (vers l'avant, vers l'arrière et de côté). En revanche, ils ont classé les systèmes de FD en deux catégories : les systèmes ambiants et les dispositifs portables. Pour cette étude,

les systèmes de vision sont considérés comme étant des détecteurs de chutes ambiants puisqu'ils sont installés dans l'environnement de la personne âgée. Néanmoins, Perry et al. [7] ont basé leur étude et leur évaluation sur l'utilisation des accéléromètres seulement.

Pour notre part nous avons choisi de regrouper les capteurs permettant de détecter les chutes en trois catégories : les capteurs embarqués sur la personne, les capteurs ambiants et les systèmes basés sur la vision. En effet, cette classification permet une meilleure compréhension des systèmes liés aux chutes. Nous l'illustrons dans la Figure 1.5. L'objectif est de fournir un schéma de référence global sur les systèmes existants liés aux chutes :

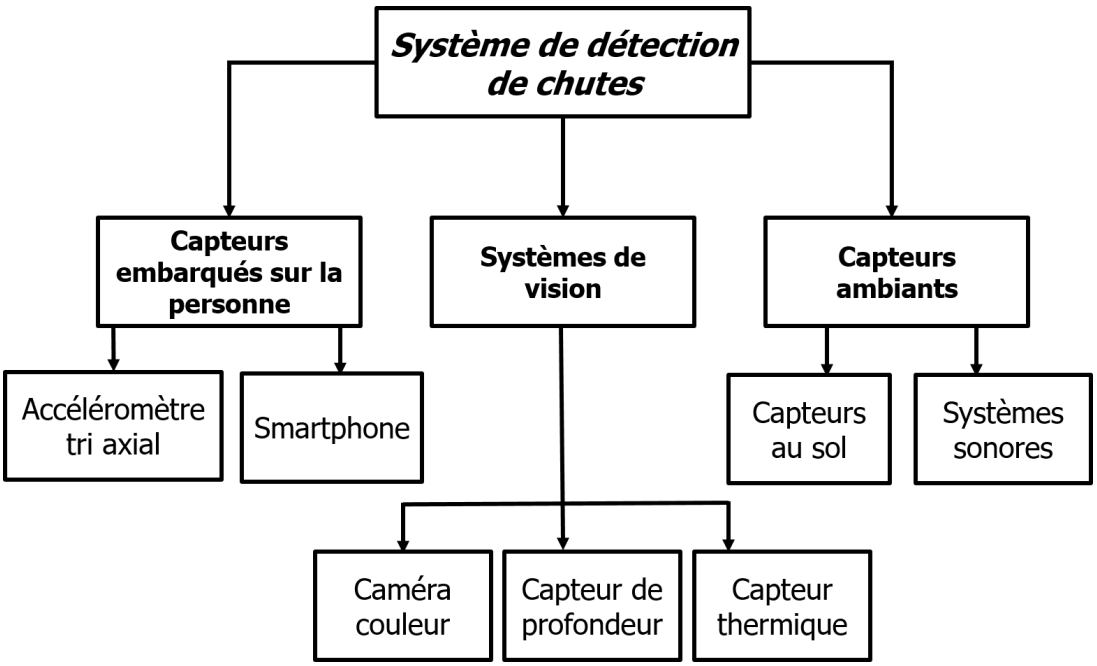


FIGURE 1.5 – Classification des systèmes de détection de chutes.

1. 2.1 Capteurs embarqués sur la personne

Un capteur embarqué sur la personne est un dispositif électronique posé sur le corps de la personne (par exemple un capteur placé au poignet, à la poitrine ou à la taille de la personne). Ce dispositif fournit des informations sur les mouvements de la personne ainsi que sa température et sa tension, ce qui est utile pour la détection de la chute. L'idée du capteur portable est de comparer l'activité normale de la personne (s'allonger, s'asseoir et marcher) à la chute. En général, les accéléromètres et les gyroscopes sont

utilisés pour mesurer l'accélération et l'orientation de chaque partie du corps pendant différentes activités. Un capteur portable est un dispositif autonome en énergie, il possède sa propre source d'énergie (pile ou batterie). Il contient une unité de traitement et une unité de communication. Il transforme les grandeurs physiques observées (température, mouvement, etc.) en données numériques compréhensibles par l'unité de traitement. Cette dernière est l'unité principale du système portable. Elle sert à gérer les périphériques, à programmer l'ordonnancement des tâches, à traiter les données et à gérer l'énergie (mise en veille et réveil du capteur selon la situation). Elle traite les données soit en local, ce qui nécessite de l'espace en mémoire, soit sur un serveur. L'unité de communication est responsable de la transmission de données au serveur via une technologie sans fil comme le WiFi ou le Bluetooth. Concernant la consommation d'énergie, la technologie WiFi consomme beaucoup d'énergie, contrairement au standard Bluetooth [8].

La Figure 1.6 décrit le passage d'information d'un système basé sur des dispositifs portables, un smartphone et un serveur.

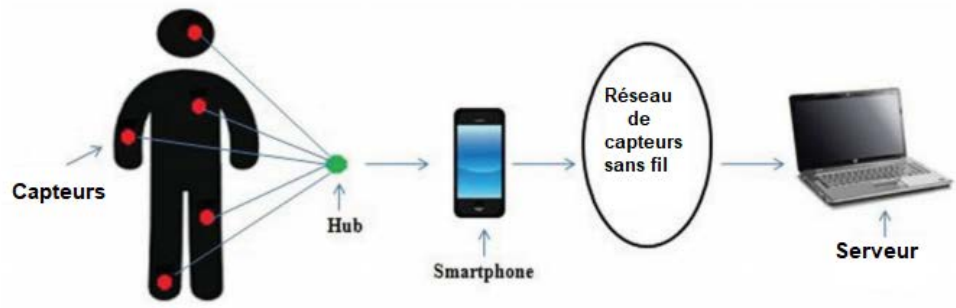


FIGURE 1.6 – Système de détection des chutes basé sur les capteurs portables [9].

Kaewkannate et Kim [10] présentent une étude comparative entre quatre dispositifs potables de type bracelets disponibles sur le marché. Ils montrent la différence de caractéristiques et de coût. Ils concluent que la consommation d'énergie de ces dispositifs dépend de la configuration du dispositif, du type de capteurs intégrés et des technologies de communication utilisées.

L'un des produits commercialisés actuellement pour détecter les chutes est la montre d'Apple. Elle analyse la trajectoire du poignet et son accélération à l'aide des accéléromètres et des gyroscopes intégrés. S'il n'y a pas de mouvement dans les 60 secondes qui suivent la détection de la chute, le service d'appel d'urgence peut être automatiquement contacté. Cependant, les systèmes portables de détection nécessitent le port de l'appareil

et c'est difficile de les mettre en permanence à la maison ou dans une chambre d'EHPAD (Établissement d'Hébergement pour Personnes Âgées Dépendantes).

Ces appareils sont simples et peu coûteux, mais ils exigent une acquisition de données de haute qualité, où les différentes informations doivent être collectées en continu. Toutefois, ces technologies présentent plusieurs limites qui rendent difficile le suivi de la personne en continue. En effet, ces limites sont notamment liées à la mémoire de stockage, la puissance de calcul et la consommation d'énergie. Pour certains cas, ces dispositifs nécessitent aussi des positions bien spécifiques (par exemple, placé au niveau de la poitrine, de la taille, des cuisses, des bras, des jambes, etc.). En conséquence, certaines personnes ne sont pas motivées pour les mettre puisque ces appareils perturbent leur vie normale.

1. 2.2 Capteurs ambiants

Pour éviter les problèmes des capteurs portables, les chercheurs ont développé une autre solution de surveillance et de détection des chutes. Ce type de systèmes repose sur des capteurs ambiants, installés directement chez la personne âgée ou dans les maisons de retraite. Les données recueillies à partir des capteurs sont envoyées à un ordinateur pour les traiter et déclencher des alarmes dans les cas d'urgence. Ces capteurs peuvent être des capteurs de pression et de vibration au sol, des capteurs infrarouges muraux ou placés au plafond ou des capteurs acoustiques.

1. 2.2.1 Système audio

Généralement les systèmes audios comprennent un réseau de microphones situés dans une pièce ou dans un espace relativement étroit et silencieux. Il est limité aux applications intérieures sans d'autres sources de bruit. Lorsqu'un bruit est détecté, le système audio amplifie automatiquement le signal. Puis, il identifie si le bruit est causé par la personne âgée ou pas. Popescu et al. [11] proposent des capteurs acoustiques pour détecter une chute. La chute est reconnue par le volume de bruit émis par deux microphones placés dans la chambre. Ce système fournit un taux de *Rappel* de 0.7 (éq. 4).
Li et al. [12] ont utilisé aussi un réseau de microphones circulaires pour recueillir les signaux sonores provenant de la personne afin de les comparer avec les signaux de la chute. Cependant, le principal inconvénient de ce type de système est son emplacement dans un environnement de vie quotidienne qui contient souvent du bruit comme les pleurs des enfants, le bruit des objets ou des animaux. En outre, le son dépend également de

chaque patient, du type d'action qu'il réalise ou de sa position. Par conséquent, ce système n'est pas assez fiable pour être appliqué dans la vie réelle.

1. 2.2.2 Capteurs au sol

Alwan et al. [13] ont développé un système de détection des chutes qui évalue les vibrations du sol à l'aide de capteurs piézoélectriques intégrés à la surface du sol. Ariani et al. [14] détectent les chutes à l'aide de capteurs sans fil à double technologie, notamment des tapis de pression et des capteurs infrarouges. Ce système a une précision de 0.89, une spécificité de 0.77 et une sensibilité de 1. Il différencie les vibrations de la chute d'une personne et de celle de la chute d'un autre objet. En revanche, ce système présente certains inconvénients. En effet, l'installation de ce dispositif implique un coût important en plus du prix du matériel. De plus, il produit fréquemment des faux négatifs, par exemple une chute sur le lit n'est pas souvent détectée. Afin de réduire le nombre de faux négatifs et d'élargir le champ de détection, certains travaux de recherche ont proposé d'intégrer des informations provenant d'autres capteurs installés dans la même pièce. Prenons l'exemple de Tzeng et al. [15], ils ont utilisé un capteur de pression au sol pour détecter un choc sur le sol et une caméra infrarouge pour identifier les activités de la personne.

Contrairement aux autres techniques, les problèmes de confidentialité sont éliminés en utilisant ces capteurs ambiants. Mais, ce type de capteurs est sensible au bruit. Il est facilement perturbé par l'environnement et coûteux.

Le développement des technologies de détection des chutes suit une progression notable. Par ailleurs, plusieurs idées d'utilisation de capteurs ambiantes restent restreintes à la recherche scientifique sans avoir recours à la commercialisation [16].

1. 2.3 Systèmes de vision

Pour améliorer la solution basée sur les capteurs ambiants et éviter les problèmes de dépendance des capteurs portables, des chercheurs ont développé des solutions basées sur la vision. Le système de vision extrait l'information des images brutes pour les traiter, comme illustré dans la Figure 1.7.

Cucchiara et al. [17] ont proposé un système basé sur une seule caméra couleur. Ils appliquent une technique de soustraction du fond pour extraire la silhouette de la personne dans la scène. Parfois, ce système ne peut pas distinguer la position accroupie de la position debout. Pour diminuer le nombre de fausses alarmes, Thome et al. [18] ont

FIGURE 1.7 – Système de détection des chutes basé sur la vision [9].

développé un système de détection de chutes multi-vues qui contient deux caméras. Le résultat atteint une *Précision* de 97,08%. Foroughi et al. [19] détectent la silhouette de la personne, qui existe dans la scène, en soustrayant l'arrière plan des images couleurs. La silhouette segmentée est entourée par une ellipse. La chute est détectée en fonction de la vitesse et l'accélération de cette ellipse. L'expérimentation a abouti à une *Précision* de 88,08%. Cependant, l'utilisation des images couleur dans la vie quotidienne ne préserve pas l'anonymat de la personne. Ainsi, la personne âgée peut refuser l'utilisation de ces systèmes.

Pour éviter toute atteinte à la vie privée, les chercheurs ont proposé d'autres types de capteurs (capteur de profondeur et capteur thermique). Les capteurs de profondeur, tels que la caméra Kinect, ont été exploités pour identifier et classer les activités humaines de la vie quotidienne. Grâce au proche infrarouge, la caméra de profondeur est indépendante de l'éclairage de la pièce où elle est installée. Elle fonctionne dans une pièce sombre ou lorsque les conditions d'éclairage changent de façon significative puisque certaines chutes peuvent être causées par la faible luminosité ou pendant la nuit. Dans l'image de profondeur, la valeur de chaque pixel représente l'information de profondeur au lieu de l'information traditionnelle de couleur. La valeur de profondeur est la distance entre l'objet et le centre optique de la caméra. Après calibration cette image en profondeur peut être transformée en information 3D pour laquelle chaque pixel de l'image est associé à un point dans le monde réel. C. Lee et al. [20] présentent un système de détection des chutes basé sur une caméra de profondeur. Cette dernière fournit des données du squelette détecté dans la scène. Ils ont choisi le milieu des hanches comme région d'intérêt (Region Of Interest ROI) pour suivre la personne. Deux conditions sont à vérifier pour valider une chute. La première condition consiste à vérifier la distance de ce centre par rapport au

sol. La deuxième condition compare la vitesse verticale de la ROI par rapport à un seuil prédéfini. Cette méthode détecte toutes les chutes dans les conditions optimales avec un taux de *Spécificité* et une *Précision* allant jusqu'à 90% et 95% respectivement. Le et al. [21] présentent un système de détection de chute basé sur le capteur Kinect. Tout d'abord, ils calculent le plan de sol de la pièce. Ensuite, ils extraient les coordonnées de la tête et de la colonne vertébrale grâce à un outil d'extraction de squelette proposé par le SDK Kinect. Enfin, ils calculent la distance de ces derniers par rapport au sol pour détecter les chutes. Ils ont obtenu une *Précision* de 83,56%, une *Précision*sensibilité de 91,12% et une *Précision*spécificité de 76%. Puis, ils ont enlevé toute image ambiguë de la base de données. Les résultats atteignent, au deuxième essai, une *Précision* de 91%, une *Sensibilité* de 100% et une *Spécificité* de 82%. Les systèmes de détection de chutes existants, qui utilisent les capteurs de profondeur, ont plusieurs avantages utiles dans le domaine de la reconnaissance d'objets par rapport aux caméras couleurs. Ils sont utiles pour résoudre le problème de changement d'échelle et pour simplifier les tâches de soustraction de l'arrière-plan et de détection du sol. Vu qu'au niveau des images couleurs, l'ombre réduit considérablement la qualité de la soustraction de l'arrière-plan. Mais, certains systèmes ont une précision faible quand la personne est dans une situation proche d'une chute.

Des solutions basées sur la vision thermique ont été explorées. L'image thermique est exploitée pour atténuer certains problèmes liés aux images couleurs [22]. Dans une image thermique composée d'un objet dans une scène, la structure de cet objet peut être facilement extraite de l'arrière-plan, quelles que soient les conditions d'éclairage et les couleurs de surfaces de l'arrière-plan, car les températures du corps humain et de l'arrière-plan sont différentes dans la plupart des situations. Par conséquent, les capteurs thermiques peuvent détecter le corps humain à l'extérieur comme à l'intérieur, de jour comme de nuit, quelles que soient les mauvaises conditions d'éclairage ou la position du corps. De plus, le prix d'une caméra thermique a considérablement baissé avec le développement de la technologie infrarouge. Quelques travaux ont utilisé des capteurs thermiques afin de protèger la vie privée des personnes âgées [23]. Quero et al. [24] ont intégré une caméra thermique à basse résolution (32×31) dans leur système. Ils ont basé la détection des chutes sur un apprentissage profond. Les résultats montrent une performance encourageante en cas d'une détection unique, avec un taux de *Précision* allant jusqu'à 92%, mais une réduction de 10% au niveau de cette métrique dans le cas d'une détection multiple. Dans le même contexte, [25] et [26] ont bien identifié les différentes activités humaines de la vie quotidienne en entraînant plusieurs réseaux de neurones, ce

qui permet d'obtenir des meilleures performances de détection de chute. Vadivelu et al. [27] sont parmi les chercheurs qui ont exploité ces capteurs pour détecter les chutes. Ils ont proposé une méthode simple basée sur le flux optique. Ils ont réussi à atteindre une *Justesse* égale à 99%. Néanmoins, ce capteur a un certain nombre d'inconvénients comme la confusion avec d'autres régions chaudes et la faible résolution.

1. 3 Systèmes de prévention de la fragilité de la personne âgée

La prévention des chutes (également appelée détection des chutes avant impact) est une autre stratégie pour détecter les chutes [28]. Comme indiqué dans le travail de Chaccour et al. [29], les technologies liées aux chutes peuvent être divisées en deux catégories : les technologies basées sur la détection des chutes ou la prévention des chutes. Hu et al. [30] ont mené une étude sur les technologies actuelles de prévention des chutes sur multiples aspects, notamment les capteurs, les indices, les méthodes de détection des chutes, les types de chutes ainsi que les performances des systèmes de prévention des chutes. Les auteurs ont signalé quelques limites de ces systèmes. L'absence des chutes réelles est la principale limite. Le travail de Kosse et al. [31] a principalement présenté une recherche exhaustive sur les technologies de détection pour la prévention des chutes chez les patients gériatriques. Quatre questions spécifiques, dont les interventions de prévention des chutes, l'efficacité des systèmes de prévention des chutes, le taux de fausses alertes et l'expérience des utilisateurs, ont été abordées et discutées en détail. Les résultats ont montré qu'il n'y avait aucune preuve que les technologies de capteurs actuelles pour prévenir les chutes de personnes dans un environnement de soins intérieurs réduiraient les taux de chute. Seule une étude parmi les 12 articles sélectionnés dans ce travail a fait état d'un taux de fausses alertes allant jusqu'à 16%. Cependant, ce taux est trop élevé. Par conséquent, des méthodes de détection efficaces doivent être ciblées pour que l'intervention soit couronnée de succès. Oladele et al. [32] ont également présenté une étude complète sur la prévention des chutes, qui visait à clarifier l'utilité de ces technologies. Le système de détection des chutes détecte généralement l'état du corps pour déclencher une alarme au cas d'urgence, tandis que le système de prévention des chutes extrait des informations sur la posture de la personne pour obtenir une alerte précoce en cas de chute. Cependant, les systèmes de détection et de prévention des chutes utilisent toujours des accéléromètres, des gyroscopes, des capteurs de pression, de la vision et des microphones, pour déterminer les chutes ou

les risques de chute. En d'autres termes, ils utilisent généralement les mêmes appareils, mais ils ont des objectifs finaux différents [28].

De nombreuses études utilisent un système de vision pour suivre les trajectoires de la tête, les changements de forme du corps ou la posture de la personne surveillée afin de détecter ou de prévenir les chutes [28]. Ces systèmes de capture peuvent être une ou plusieurs caméras RGB [33, 34, 35, 36], de profondeur (Kinect) [37, 38, 39, 40, 41, 42], thermiques [23, 27, 43], ou même une combinaison de plusieurs caméras. La méthode basée sur la vision la plus simple et la plus courante applique une seule caméra, comme celle utilisée par de Miguel et al. [33], qui utilise l'algorithme k-Nearest Neighbours (kNN) pour analyser le comportement de la silhouette dans le temps. Multiple caméras sont utilisées dans plusieurs systèmes afin de couvrir une large zone de détection, par exemple, Fan et al [35] ont appliqué 8 caméras montées dans une pièce. Les capteurs de profondeur sont utilisés pour calculer la distance entre la personne et le sol afin d'améliorer les performances du système de détection des chutes. Par exemple, les deux travaux de Zhao et al [37] et Li et al [38] ont suivi les principales articulations du corps en utilisant une caméra de profondeur pour détecter ou prévenir les chutes. En outre, les capteurs thermiques sont également utilisés dans les recherches sur les chutes, qui ont une justesse allant jusqu'à 99,7% [44].

La Table 1.3 représente plusieurs études sur les systèmes de détection et de prévention des chutes basés sur les systèmes de vision.

1. 3.1 Conclusion

Pour obtenir une détection automatique et efficace, nous avons besoin d'un système précis, peu coûteux et convivial. Les capteurs cinématiques sont économiques. Mais, ils présentent certains inconvénients. Ce type de capteur doit être porté jour et nuit. De plus, l'option des boutons est également inefficace lorsqu'une personne devient inconsciente. Enfin, une source d'énergie est nécessaire pour les faire fonctionner. Les capteurs ambiants et acoustiques sont fixés dans l'environnement de la personne, ce qui évite les problèmes des capteurs portables. Néanmoins, ils sont sensibles aux bruits. Ils produisent un taux important de faux négatifs et leurs coûts d'installation et de maintenance sont élevés. Les systèmes basés sur la vision sont de plus en plus utilisés pour détecter les chutes. Ils ne sont pas affectés par le bruit, contrairement aux dispositifs sonores. Ils se fixent dans la pièce à surveiller et ils peuvent être installés facilement. De plus, ils ne sont pas chers. Compte tenu de la protection de la vie privée, les systèmes de vision, en dehors des

Type de capteur	Détection de la chute			Prévention de la chute		
	Article et l'année	**Capteur et emplacement**	**Performance**	**Article et l'année**	**Capteur et emplacement**	**Performance**
Vidéo	Miguel et al. (2017) [33]	Camera (hauteur : 2-2.25m)	Justesse : 96.9% Spécificité : 97.6%	Kutchka &al (2016) [34]	Camera embarqué au mur	Pas mentionner
	Fan et al. (2017) [35]	8 cameras	Justesse : 95.2%	Li et al. (2017) [36]	Camera RGB et chaussures avec laser	Réduire le risque des chutes
Capteur de profondeur	Zhao et al. (2018) [37]	Kinect V2 / Orbbec Astra	Justesse : 97.1% Spécificité : 100%	Li et al. (2018) [38]	Kinect 2.0	Spécificité : 81.3%
	Kong et al. (2017) [39]	Kinect	Justesse : 97.1% Spécificité : 100%	Xu et al. (2017) [40]	Plusieurs Kinect	Justesse: 91.7%
	Akagunduz et al. (2017) [41]	Vidéos de profondeur	Justesse : 89.63%	Dubois et al. (2014) [42]	Kinect	Précision suffisante pour l'utilisation au milieu réel
Capteur thermique	Rafferty et al. (2016) [23]	Caméra thermique au plafond	Justesse : 68%	Song et al. (2017) [43]	Caméra thermique	Justesse : 99.7%
	Vadivalu et al. (2016) [27]	Caméra thermique	Justesse : 99.61%			

TABLE 1.3 – Les techniques de détection et de prévention des chutes [28]

caméras couleurs, fournissent un moyen discret et non intrusif d'observation des personnes et leurs activités, comme le montre la Figure 1.8.

Systèmes de détection de chutes

Capteurs portables	Capteurs ambiants	Systèmes de vision
Avantages : • Prix bas • Installation et configuration facile • Préserve l'anonymat	**Avantages :** • Préserve l'anonymat • Système non intrusif • Discret	**Avantages :** • Détection de plusieurs événements au mêmes temps • Préserve l'anonymat • Taux de détection élevé • Prix bas
Inconvénients : • Facile à se déconnecter • Taux de fausses alarmes moyen à élevé • Capteur intrusif	**Inconvénients :** • Prix élevé • Installation moyenne à difficile • Sensibles de tout type du bruit dans l'environnement • Taux de fausses alarmes élevé	**Inconvénient :** • Caméra couleur ne protège pas la vie privée de la personne • Difficulté de configuration

FIGURE 1.8 – Comparaison des systèmes de détection des chutes.

Chapitre 2

Description du contexte général de la thèse

Dans le premier chapitre, nous avons décrit des systèmes de détection des chutes basés sur différents types de capteurs. Ces derniers sont classés en trois catégories : capteurs embarqués sur la personne, capteurs ambiants et systèmes de vision. Dans ce chapitre nous présentons dans un premier temps les capteurs de vision, plus particulièrement les capteurs de profondeur et thermique, utilisés dans notre dispositif. Dans un second temps, nous expliquons l'étape de calibration qui permet de combiner les deux informations. Enfin, nous décrivons les bases de données utilisées dans cette thématique.

2. 1 Projet PRuDENCE

Les séquelles physiques et psychologiques liées aux chutes chez les personnes âgées sont associées à des coûts importants. Prenons l'exemple de la France, ces séquelles entrainent des coûts de l'ordre de un à deux milliards d'euros par an en soins de santé directs. Par contre aux États-Unis, les coûts sont estimés à 19,2 milliards de dollars pour les blessures mortelles et non mortelles. Concernant la répartition des coûts de cette dernière, on note que 63 % (12 milliards de dollars) sont relatifs aux hospitalisations, 21 % (4 milliards de dollars) correspondent aux visites aux services d'urgence et 16 % (3 milliards de dollars) sont liés aux traitements en milieu ambulatoire [45].

La mise en œuvre de stratégies d'intervention efficaces pourrait réduire les coûts des soins de santé liés à ces blessures.

La détection et la prévention de la chute permettent de diminuer les coûts humains et économiques de manière considérable. En effet, les conséquences des chutes s'aggravent avec le temps passé au sol, ces conséquences pouvant même mener au décès de l'individu. À court terme, il est donc impératif de **détecter les chutes** de manière fiable afin de pouvoir porter secours à la personne le plus rapidement possible. À plus long terme il est

également très important de prévenir les chutes. Une des possibilités de prévention est le **suivi à long terme de l'activité** de la personne âgée. L'idée est alors de détecter au plus tôt la fragilisation de la personne et d'y remédier par des gestes adéquats pour objectifs de maintenir au maximum les personnes au domicile, en autonomie et donc de prédire et de retarder l'apparition de la première chute ou de diminuer le nombre d'occurrences de ces chutes [46].

Dans ce contexte, le projet PRuDENCE (PRévention et DEtectioN des ChutEs) a été sélectionné dans le cadre des appels à projets génériques de l'ANR en 2016. Il a comme partenaire la société NeoTec-Vision (vision industrielle), le LTSI Université de Rennes 1 (traitement d'images et analyse de données), l'ECAM Rennes - Louis de Broglie (traitement d'images), le Living Lab ActivAgeing (LL2A) de l'Université de Technologie de Troyes (évaluation et acceptabilité des soulutions) et l'équipe d'accueil EA2694 Santé Publique/épidémiologie et qualité des soins de l'Université de Lille (aspects réglementaires des données et des solutions).

L'objectif du projet PRuDENCE était de proposer un nouveau dispositif à bas coût à base de capteurs de profondeurs et/ou thermiques et leurs traitements associés permettant de prévenir le risque de chute par l'analyse de l'activité des individus. Ce dispositif doit permettre de maintenir les personnes en autonomie à leur domicile et en leur assurant ainsi une meilleure qualité de vie. Ceci implique d'ailleurs une participation active de la personne âgée qui doit se sentir considérée comme partenaire de sa propre santé.
La prise en compte de la personne âgée est intervenue depuis la genèse du projet PRuDENCE. En effet, la société NeoTec-Vision, lorsqu'elle a voulu développer une solution de détection de personnes au sol, a demandé une enquête d'acceptabilité psychosociale menée avec le LAUREPS (Laboratoire Armoricain Universitaire de Recherche En Psychologie Sociale - Université de Rennes 2). Cette enquête a démontré qu'un tel dispositif se devait avant tout d'être passif et être basé sur l'idée de "Sécuriser sans surveiller" (et donc de **préserver** complètement l'**anonymat** des individus observés). Ceci a conduit, avec une collaboration scientifique entre NeoTec Vision, LTSI, Université de Rennes 1 et l'ECAM Rennes -Louis de Broglie, au développement d'un premier prototype basé sur une caméra de profondeur associée à une caméra thermique qui permet une détection de présence au sol fiable, rapide, sans contact et non intrusive, marchant jour et nuit, et qui peut être raccordée aux systèmes d'appels malades existants [47]. Ce prototype a subi une première phase de test dans deux EHPAD. Ce test a permis de montrer certaines limites de la solution pour la détection de chutes (de présence au sol) avec en particulier

la gestion des occultations. De plus, ces tests ont également permis de démontrer que cette solution initiale, limitée à la seule détection de présence au sol, pouvait être étendue potentiellement à la localisation et au suivi de l'activité d'une personne.

En plus d'une étude sur une solution de remplacement du couple de caméras de profondeur/thermique par une solution à plus bas coût avec une paire stéréoscopique de caméras thermiques, le projet PRuDENCE s'est monté autour de cette idée avec un axe sur une solution de **suivi temporel combinant les modalités profondeur/thermique** et un axe sur **l'analyse des scénarii de vie** de proposer des outils de diagnostic précoce pour des difficultés de locomotion afin de faire de la prévention. Ces différentes solutions seront développées et évaluées en collaboration avec LL2A qui d'une part dispose de solutions de recueil de positions dans l'espace mais également réalise des démarches itératives de co-conception afin d'assurer le développement d'un dispositif en adéquation avec les besoins des utilisateurs et évalue la pertinence (acceptabilité, utilisabilité, apports médico-sociaux) des solutions.

Le travail de cette thèse se situe dans l'axe "suivi temporel combinant les modalités cartes de profondeurs images/thermique". Le suivi d'une personne dans une séquence d'images est un cas particulier de suivi d'objets qui représente une thématique ancienne mais toujours difficile et d'actualité en analyse d'images. Il est particulièrement complexe lorsque l'objet est hautement déformable comme c'est le cas pour une personne. Il se complique encore lorsque l'on doit considérer d'inévitables occultations partielles ou totales, par exemple lorsqu'une partie de la personne est masquée par un meuble ou que le mobilier lui même peut être déplacé (chaise,..).

Les objectifs du travail de Thèse sont : 1) d'**optimiser le suivi d'une personne** dans une scène complexe avec des occultations possibles à partir d'une information provenant de cartes de profondeurs issues d'un capteur bas-coût (Kinect) et d'images thermiques issues de capteurs infrarouge basse-résolution (80×60 pixels) ; 2) d'**estimer des paramètres permettant d'identifier des postures** telles que la position allongée, la position assise, la chute, etc. L'estimation de ces paramètres et leur suivi temporel est indispensable pour l'axe *analyse des scénarii de vie*. Pour cela les activités et les postures à estimer devront se rapprocher le plus possible des situations des personnes fragiles en autonomie à domicile.

Dans la suite de ce chapitre nous allons présenter, dans un premier temps, les caractéristiques des deux capteurs (profondeur et thermique) utilisés par le dispositif surveillance, ainsi que le principe de calibration utilisé pour permettre la fusion des images issues de ces deux modalités. Dans un second temps, nous allons présenter les quelques bases de

données publiques disponibles et surtout les bases de données que nous avons créées pour mettre en œuvre nos algorithmes.

2. 2 Types de capteurs

La solution matérielle proposée dans le cadre du projet PRuDENCE doit être à bas-coût, donc basée sur des capteurs existants et bon marché. Dans cette section, nous allons présenter les caractéristiques des deux capteurs (profondeur et thermique) choisis pour le projet, ainsi que l'étape de calibration nécessaire à la fusion des deux images.

2. 2.1 Capteur de profondeur

2. 2.1.1 Principes d'un capteur de profondeur

Il existe une très grande gamme de capteurs de profondeur. Ces capteurs peuvent être classés selon la technologie de création des cartes de profondeur en deux catégories :

— Motif de lumière structurée. La méthode de la lumière structurée a pris un intérêt important ces dernières années en raison de sa précision pour le calcul de distances. Son principe est le suivant : un projecteur de lumière structurée remplace l'une des caméras de la méthode de stéréo vision. Ce dernier émet certains pattern de lumière, notamment les motifs ponctuels, linéaires et codés vers les objets. La réflexion du motif est captée par la caméra. L'information de profondeur est calculée sur la base de la triangulation [48].

— Principe du temps de vol (Time Of Flight TOF). Pour les capteurs de profondeur basés sur le principe du TOF, une lumière proche infrarouge modulée est émise pour éclairer les objets. L'information de profondeur est obtenue en comparant le déphasage entre la lumière émise et celle réfléchie par l'objet observé [49].

Les capteurs de profondeurs présentent plusieurs avantages par rapport aux systèmes de vision classiques. Ils sont indépendants de la couleur des objets et de l'éclairage de la pièce, chose extrêmement importante pour notre projet. Il est extrêmement difficile de reconnaitre les personnes filmées et donc ils préservent leur anonymat. L'avantage d'un capteur de profondeur est qu'il permet d'obtenir directement une carte de profondeur alors que les systèmes de stéréo-vision classiques nécessitent un post-traitement fiable pour recouvrir une information de profondeur plus ou moins précise et dépendant du contenu des images. Cet accès direct à l'information 3D rend le traitement de la détection de la

chute plus rapide. Dans la littérature, il existe une multitude de capteurs de profondeur. Pour notre étude, nous nous limitons au système de recueil de cartes de profondeur le plus répandu, et donc de plus bas coût, le système Kinect proposé par Microsoft dans le cadre de leur console de jeux. Dans la suite de ce chapitre, nous allons détailler les caractéristiques de ce système.

2. 2.1.2 La caméra Kinect

La caméra Kinect est un dispositif du suivi des mouvements humains qui a été développé pour la console Xbox 360 en novembre 2010 [50]. La première version est connue sous le nom de Kinect V1 (Figure 2.1.a). En 2014, Microsoft a développé une deuxième version V2 (Figure 2.1.b) qui améliore certaines caractéristiques matérielles et logicielles de sa première version [51].

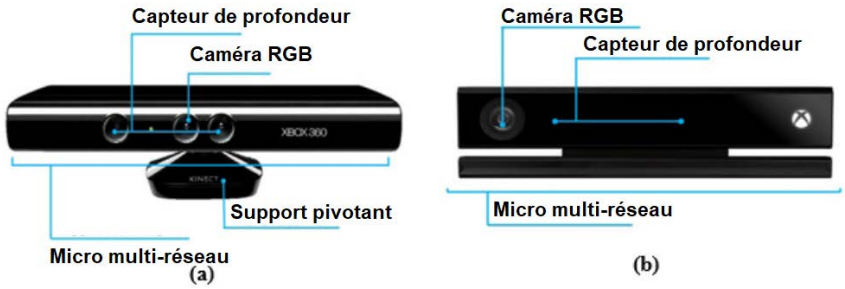


FIGURE 2.1 – Les caméras KINECT V1 (a) et KINECT V2 (b).

Les caméras Kinect sont caractérisées selon plusieurs critères dont les plus importants sont :

— **Capteurs :** une caméra RGB qui fournit des images optiques classiques en couleurs (non utilisées dans notre cas), un système de capteur de profondeur et un réseau de micros à reconnaissance vocale. Concernant le capteur de profondeur, il existe une grande différence de conception entre le système utilisé par la caméra kinect V1 qui est constitué par un couple projecteur/caméra infra-rouge et celui de la caméra kinect V2 qui utilise une caméra Time Of Flight (TOF). Le recueil de l'information de profondeur par couple projecteur/caméra pour la caméra Kinect V1 explique la présence d'**ombres** de surface non détectées dans les images fournies (cf Figure 2.2).

— **Champ de vision :** les champs de vision (Field of View FOV) verticaux (v) et

FIGURE 2.2 – Exemple d'images obtenues avec a) une Kinect V1 et b) une Kinect V2 [52]

horizontaux (h) sont de l'ordre de (57° (h), 43°(v)) et (70° (h), 60° (v)) respectivement pour les versions V1 et V2.

— **Portée du capteur :** la portée des capteurs est de l'ordre de 0,8 m à 3,5 m et de 0,5 m à 4,5 m respectivement pour la version V1 et la version V2.

— **Flux de données :** la résolution des images couleur et de profondeur est égale à 640×480 pixels pour la version V1. Une amélioration significative dans la résolution de l'image RGB a été apportée à la version V2 (1920 × 1080 pixels, mais nous n'utilisons pas ce capteur). Par contre pour l'image de profondeur, la résolution a été abaissée à 512×424 pixels. En ce qui concerne les systèmes audio, le réseau de la version V1 est composé de quatre microphones pour assurer une communication en direct. Celui-ci permet de traiter des signaux audio jusqu'à 16 bits à une fréquence d'échantillonnage de 16 kHz. Concernant la version V2, avec le même nombre de microphones, la fréquence d'échantillonnage atteint 48 kHz. Ce système d'audio permet de supprimer les échos et reconnaitre plusieurs langues.

— **Moteur d'inclinaison :** un support pivotant n'existe que sur la caméra Kinect V1 pour permettre le suivi de déplacement des personnes en inclinant automatiquement le capteur jusqu'à 27° en haut ou en bas.

2. 2.1.3 Les avantages des systèmes Kinect

De nombreux chercheurs font de Kinect leur premier choix de capteur de profondeur car elle s'avère être la première caméra qui combine tous les avantages suivants :

— **Combinaison de plusieurs flux :** la caméra Kinect combine plusieurs types d'in-

Caractéristiques	Description
Dimension	27.94 x 6.35 x 3.81 cm
Résolution	640 (h) x 480 (v) pixels
Champs de vision (FOV Field Of View)	57° (h) x 43° (v)
Moteur d'inclinaison	Automatique jusqu'à 27°
Fréquence	30 Hz
Prix	170$

TABLE 2.1 – Spécifications de Kinect V1

formation RGB, de profondeur, squelette, infrarouge et audio dans un seul système ce qui permet de l'utiliser dans plusieurs domaines. Les images de type RGB-D (couleur et profondeur) peuvent améliorer considérablement les performances des classifieurs d'objets.

— **Dédiée pour les applications d'intérieur :** la caméra Kinect a été créée pour le pilotage de jeux vidéos donc adaptée à une utilisation en intérieur. Elle peut être utilisée directement, sans adaptation au niveau des systèmes de détection des chutes dédiés aux hôpitaux et aux maisons de retraite.

— **Sécurité :** le laser infrarouge (IR) Kinect V1 est doté d'une lumière infrarouge de type 2 qui n'est pas dangereuse pour l'homme, contrairement à la plupart des scanners laser. Ces derniers fonctionnent avec un laser de type 1 qui est néfaste pour les yeux en absence de protection. Ainsi, Kinect garantit la sécurité de la personne.

— **Facile à utiliser :** tous les systèmes à bases de caméras sont des systèmes passifs ne nécessitant aucune interaction avec la personne âgée (contrairement aux systèmes portables).

— **Faible coût :** ce type de capteur est disponible partout dans le monde et il coûte une centaine d'euros, y compris l'adaptateur et le kit de développement logiciel.

— **Préserve l'anonymat :** dans notre étude, seules les images de profondeur sont utilisées. La personne suivie n'est pas reconnaissable sur de telles images de profondeur. Ceci permet de mieux protéger son anonymat par comparaison aux images RGB classiques.

— **Indépendant aux conditions d'éclairage :** le capteur Kinect est capable d'extraire les cartes de profondeur dans les pièces sombres à l'aide de ces capteurs infrarouges. Il peut donc détecter les chutes même pendant la nuit.

— **Fréquence élevée :** il peut être utilisé pour les applications en temps réel grâce

à la fréquence d'images (FPS à 30 Hz)

Ces caractéristiques permettent à Kinect d'être utilisée pour la détection des mouvements en 3D du corps entier, la reconnaissance faciale et la reconnaissance vocale. Ce qui la rend très utile dans diverses applications telles que : le suivi et la reconnaissance d'objets, le suivi du squelette humain et l'analyse de l'activité, l'analyse des gestes de la main, la localisation et la cartographie 3D simultanée, la détection des urgences comme les agressions, les chutes et autres [51].

2. 2.1.4 Comparaison de la caméra Kinect avec d'autres capteurs de profondeur

Nous avons comparé la caméra Kinect V1 avec d'autres capteurs de profondeur plus récents, détaillés dans la Table 2.2. Nous constatons que la version V1 semble offrir le meilleur compromis entre une résolution et une qualité correctes et un coût faible.

En revanche, il a été noté par les utilisateurs certaines caractéristiques négatives comme l'ombre sans signal due à la parallaxe entre l'émetteur et le récepteur infrarouge et aussi certains effets inhérents aux infrarouges comme une sensibilité aux bruits infrarouges tels que le soleil et ses reflets, ainsi qu'à la présence d'objets absorbants, qui peut dégrader la qualité de l'image ou la précision de la mesure en profondeur.

2. 2.2 Capteur thermique

Les capteurs thermiques sont des capteurs qui mesurent le rayonnement infrarouge lointain qui correspond globalement au rayonnement thermique émis par les objets. Ce type de capteurs est particulièrement bien adapté à notre problématique de suivi de personnes, car les personnes, du fait de la chaleur interne, se différencient facilement de l'arrière-plan (Figure 2.3). La silhouette peut donc être facilement détectée, même durant la nuit où la probabilité de chute est élevée. Si les capteurs infrarouges ont été considérés comme chers, certains constructeurs proposent actuellement des capteurs bas-prix afin de s'adapter au marché des smartphones. Par contre, ce prix bas se fait au détriment de la qualité de l'image et surtout de sa résolution.

Pour notre étude, nous avons sélectionné un des capteurs les moins chers, la caméra Lepton 2 du fabriquant FLIR. Ce bas coût permettra de proposer une solution commerciale pratiquement viable. Les spécifications du capteur Lepton sont données dans la Table 2.3. Il est à noter la faible résolution (80×60 pixels) de l'image. Comme la Lep-

Nom du capteur	**Capteur**	**Prix**	**Technologie**	**Résolution de l'image de profondeur**	**Fréquence (FPS)**	**Année**
Azure Kinect		400 $	Principe du temps de vol (Time of flight TOF)	512x512	30	2019
Kinect V2		280$	Principe du temps de vol TOF	512x424	30	2014
Kinect V1		170$	Motif de lumière structuré (Infrared Coded Structured Light)	640x480	30	2010
Orbbec Astra (PRO)		179$	Motif de lumière structuré (Infrared Coded Structured Light)	640x480	30	2015
ASUS Xtion 2		270$	Principe du temps de vol TOF	640x480	30	2018

TABLE 2.2 – Comparaison des capteurs de profondeur.

ton a été conçue pour être intégrée à des appareils mobiles, elle est très compacte, avec une longueur maximale de 11,7 mm et un poids de seulement 0,55 g. Afin d'utiliser ce capteur nous l'avons attaché à un module intelligent intégré proposé par FLIR, appelé Purethermal I FLIR Lepton. Ce module est livré pré-configuré pour faire fonctionner le capteur thermique via un port USB UVC 1.0. Ce module permet de faire fonctionner le capteur Lepton avec des applications webcam standard telles que VLC Media Player sur PC, Linux, Mac et même Android. FLIR propose également un boitier qui protège le module et la caméra thermique, tout en permettant l'accès aux boutons, au port micro USB et à la lentille du capteur thermique (voir Figure 2.4).

Ce choix de capteur est justifié par :

— la facilité d'extraire la silhouette de la personne
— la préservation de son anonymat sur l'image
— l'utilisation de jour comme de nuit.

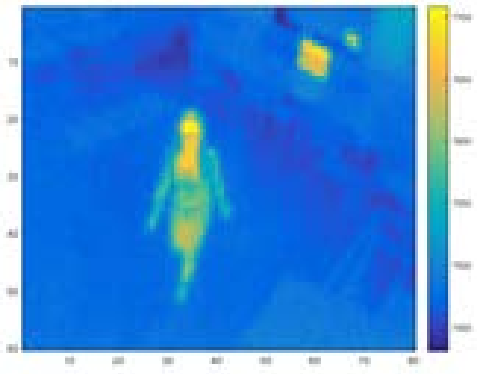


FIGURE 2.3 – Exemple d'image thermique.

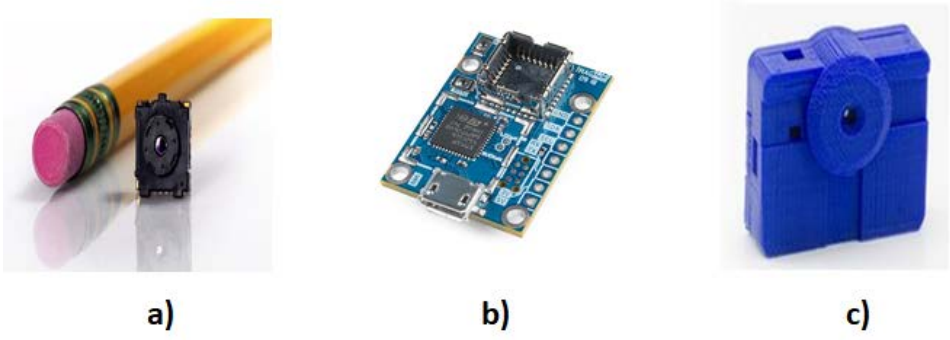


FIGURE 2.4 – a) Caméra Flir - Lepton 2.5, b) Module Purethermal 1 FLIR Lepton et c) Boitier

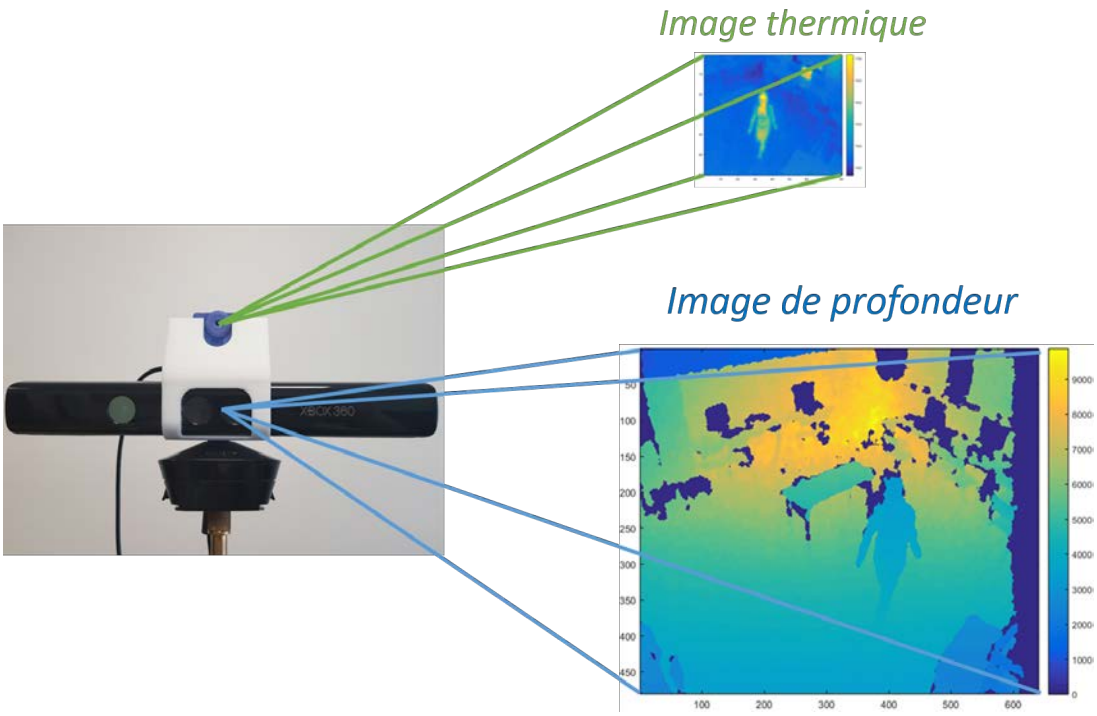


FIGURE 2.5 – Images obtenues par notre système d'acquisition.

Caractéristiques	Description
Dimension	11.8 x 12.7 x 7.2 mm
Résolution	80 (h) x 60 (v) pixels
Champs de vision (FOV)	51° (h) x 37.83° (v)
Sensibilité thermique	<50 mK
Justesse	±5° C
Fréquence	9 Hz
plage de longueurs d'onde du champs spectral	8 à 14 microns (nominale)
Prix	< 150$

TABLE 2.3 – Spécifications de Flir - Lepton 2.5.

2. 3 Association des capteurs et calibration

Nous avons fixé les capteurs l'un sur l'autre d'une façon que les axes optiques soient en parallèle. La disposition des capteurs est illustrée par la Figure 2.6 ; le capteur thermique est fixé sur le capteur de profondeur. Les deux capteurs sont donc solidaires à l'aide d'un support conçu et imprimé à l'Ecam et ont globalement le même champ de vue. Il est à noter que les deux images n'ont pas la même résolution : 640×480 pour le capteur de profondeur et 80×60 pour le capteur thermique. Par contre le rapport de résolution est exactement de 8 dans les deux directions. La Figure 2.5 illustre cette différence de résolution sur un exemple d'images obtenues par notre dispositif d'acquisition.
Nous disposons donc de deux informations complémentaires de la même vue, d'une part la chaleur émise par la personne et son environnement (ce qui va nous être utile pour estimer la silhouette de la personne) d'autre part la distance des différents objets par rapport au centre de la caméra de profondeur (ce qui va nous être utile pour localiser cette silhouette par rapport à son environnement, distance par rapport au sol, etc ..). Pour assurer cette complémentarité il faut fusionner ces informations, c'est-à-dire produire une "image" pour laquelle le contenu de chaque pixel est un vecteur valeur thermique/profondeur. Pour cela il nous faut une étape de **calibration**. Cette étape de calibration ne sera effectuée qu'une seule fois après avoir fixé les capteurs (au plafond par exemple). L'objectif de notre calibration est de faire correspondre un pixel de l'image de profondeur à son pixel équivalent dans l'image thermique. Cet objectif est assez proche de la calibration d'une paire de caméras en stéréoscopie. Dans la littérature, la calibration de deux caméras stéréo (ou plus) est souvent assurée par l'acquisition par les différents capteurs d'une même mire d'étalonnage. Généralement cette mire est de type grille de points ou échiquier classique en noir et blanc.

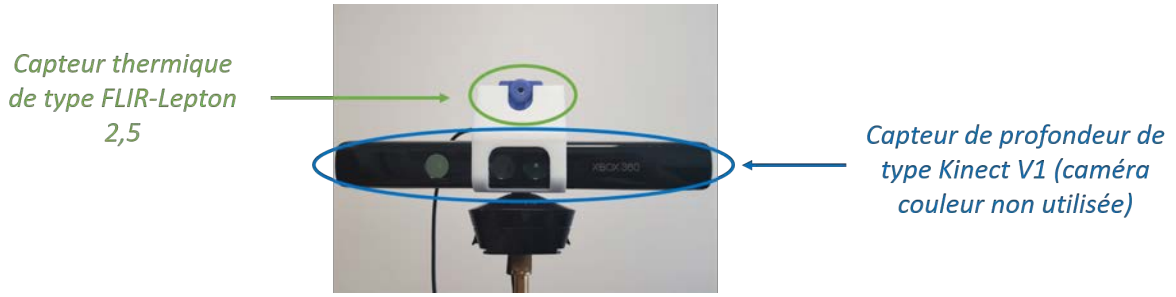


FIGURE 2.6 – Disposition des capteurs.

En revanche, l'information utile d'une mire classique ne peut être vue ni par le capteur thermique ni par celui de profondeur. Pour ces raisons, nous avons décidé de concevoir notre propre mire de calibration. Elle est constituée de plusieurs tubes de différentes hauteurs montés ensemble sur une planche. Sur chacun de ces tubes nous avons placé une résistance chauffante (voir Figure 2.7).

FIGURE 2.7 – Mire de calibration en vue de face (a) et en vue de gauche (b).

L'idée de cette mire est simple. Les tubes seront détectés par le capteur de profondeur et la chaleur émise par les résistances sera visualisée par le capteur thermique. La Figure 2.8 nous montre la mire vue par le capteur RGB (non utilisé dans notre projet) et le capteur de profondeur de la caméra Kinect, ainsi que l'image thermique obtenue par le capteur lepton.

L'étape de calibration permet d'estimer la relation entre le repère de l'image de profondeur et celui de l'image thermique en prenant en compte les coordonnées spatiales 3D des

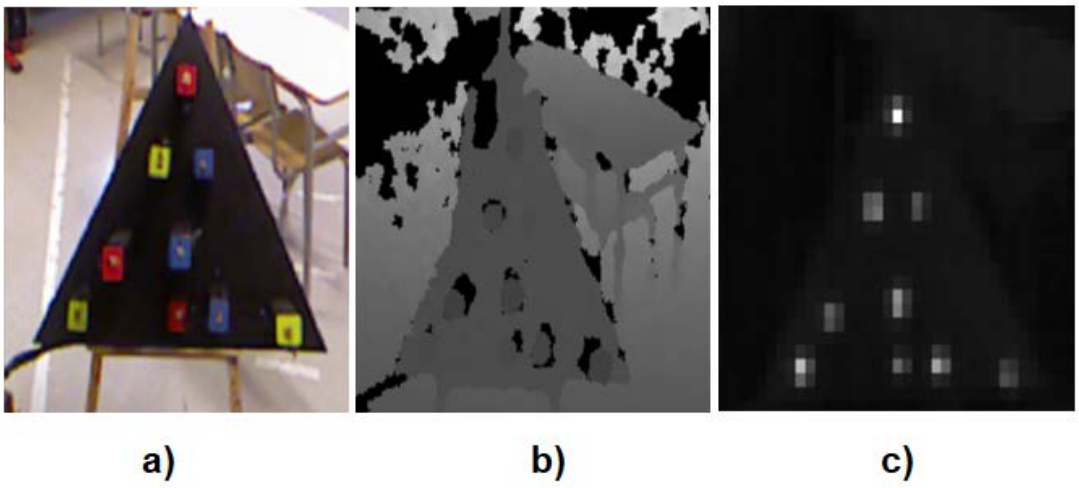


FIGURE 2.8 – La mire de calibration a) en couleur, b) sur l'image de profondeur et c) sur l'image thermique.

points remarquables de la mire. Autrement dit un pixel de l'image thermique correspond à un bloc de 8x8 pixels de l'image de profondeur.

La Figure 2.9 présente les différents repères utilisés dans notre étape de calibration. L'objectif de la calibration est de connaître la relation entre le repère de l'image thermique (u_{th}, v_{th}) et le repère de l'image de profondeur (u_d, v_d). Pour cela nous allons passer de manière indirecte par le repère 3D du capteur thermique (x_{th}, y_{th}, z_{th}) et le repère 3D du capteur de profondeur (x_d, y_d, z_d). La relation entre un repère image et un repère 3D est donnée par le modèle "pinhole" (Figure 2.10). Par souci de simplicité, nous considèrerons que le centre optique est le centre de l'image.

L'estimation de la relation entre le repère (u_{th}, v_{th}) et (u_d, v_d) nécessite alors 3 étapes (attention, les repères capteurs sont indirects - voir Figure 2.11) :

— a) Transformation des coordonnées de l'image de profondeur (u_d,v_d) en coordonnées dans le repère du capteur Kinect (x_d,y_d,z_d). Cela peut être fait analytiquement à partir des paramètres intrinsèques de la caméra de profondeur (équation 2.1) :

$$\begin{cases} x_d = (u_d - \frac{Xd_{Res}}{2})\frac{2w_d \tan(\frac{D_{HFOV}}{2})}{Xd_{Res}} \\ y_d = -(v_d - \frac{Yd_{Res}}{2})\frac{2w_d \tan(\frac{D_{VFOV}}{2})}{c} \end{cases} \tag{2.1}$$

Xd_{Res} et Yd_{Res} sont les résolutions verticales et horizontales de l'image de profon-

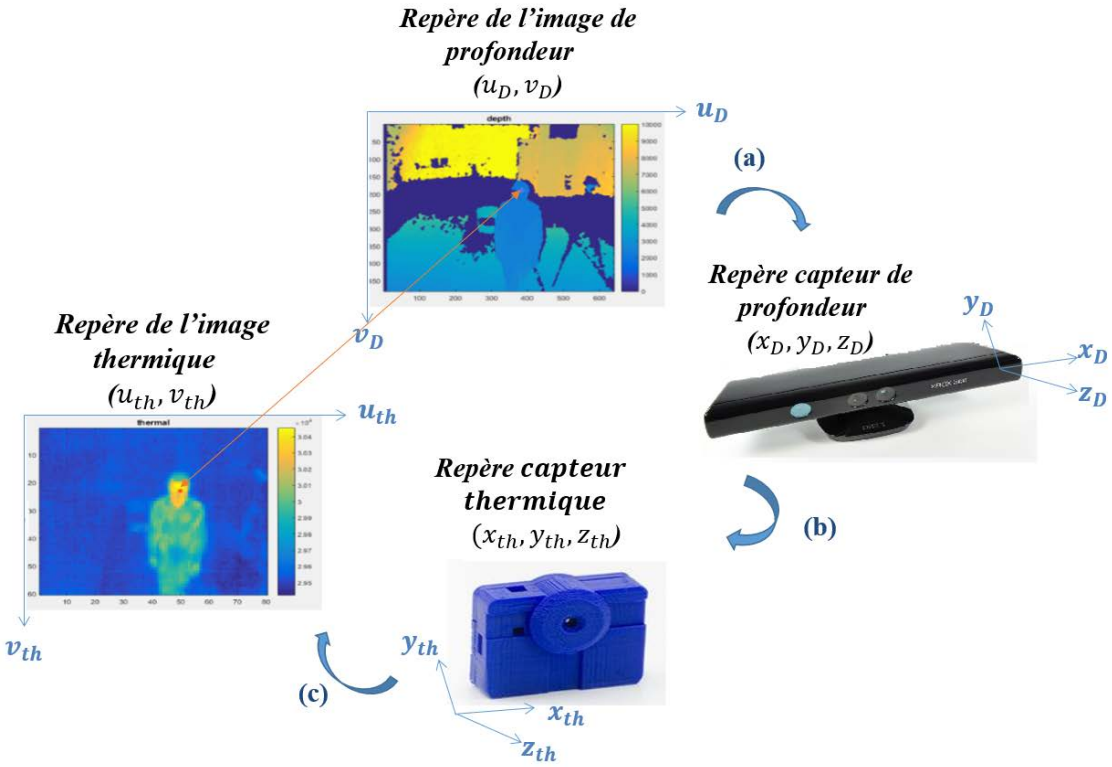


FIGURE 2.9 – Repères utilisés pour la calibration.

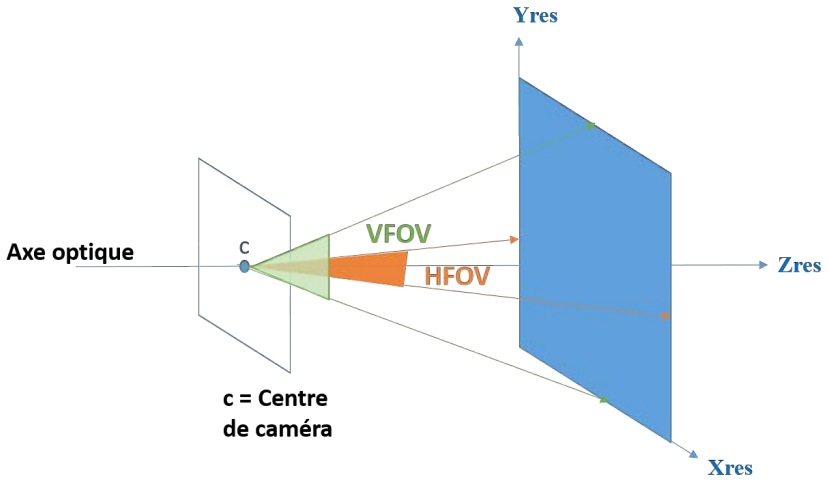


FIGURE 2.10 – Modèle géométrique pinhole.

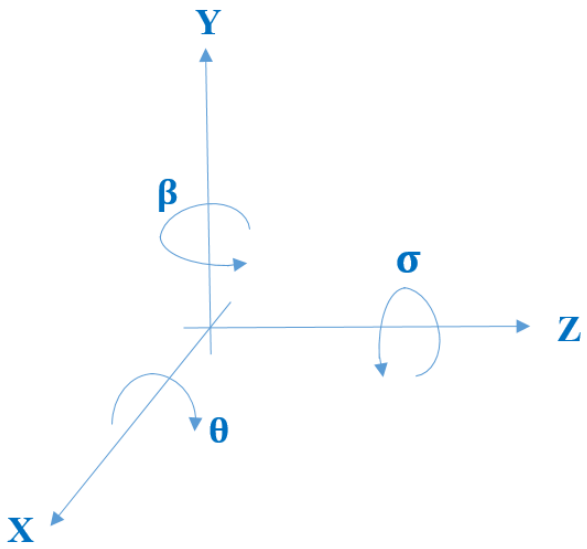


FIGURE 2.11 – Repère capteur et angles de rotation.

deur en pixels, DHFOV et DVFOV sont les ouvertures horizontale et verticale de la caméra Kinect.

— b) Estimation des coordonnées du repère Lepton-FLIR (x_{th},y_{th},z_{th}) à partir des coordonnées du repère Kinect (x_d,y_d,z_d). Le repère Kinect est l'image du capteur thermique par une transformation composée d'une rotation R et une translation T :

$$R = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} * \begin{pmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{pmatrix} \tag{2.2}$$

avec α, θ et β les angles de roulis, de tangage et de lacet (Figure 2.11).

$$T = \begin{pmatrix} d_x \\ d_y \\ d_z \end{pmatrix} \tag{2.3}$$

$$\begin{pmatrix} x_{th} \\ y_{th} \\ z_{th} \end{pmatrix} = T + R * \begin{pmatrix} x_d \\ y_d \\ z_d \end{pmatrix} \tag{2.4}$$

— c) Transformation des coordonnées dans le repère du capteur thermique (x_{th}, y_{th}, z_{th}) en coordonnées dans le repère de l'image thermique (u_{th}, v_{th}). Cette transformation s'effectue à partir des paramètres intrinsèques de la caméra thermique (voir équation 2.5) :

$$\begin{cases} u_{th} = \frac{Xth_{Res}}{2z_{th}\tan(\frac{THFOV}{2})}x_{th} + \frac{Xth_{Res}}{2} \\ v_{th} = -\frac{Yth_{Res}}{2z_{th}\tan(\frac{TVFOV}{2})}y_{th} + \frac{Yth_{Res}}{2} \end{cases} \tag{2.5}$$

Xth_{res} et Yth_{res} sont les résolutions verticales et horizontales de l'image thermique en pixels, THFOV et TVFOV sont les ouvertures horizontale et verticale de la caméra Lepton. Dans notre cas afin de simplifier la calibration, les paramètres intrinsèques sont estimés directement à partir des valeurs X_{res}, Y_{res}, HFOV et VFOV données par les constructeurs des capteurs (Tables 2.3 et 2.1). L'objectif de la calibration est alors d'estimer les 3 paramètres de translation (d_x, d_y, d_z) ainsi que les 3 angles de rotation (α, θ et β). Pour cela nous avons pris 15 paires d'images de la mire et nous avons mis en correspondance manuellement 100 pixels dans ces paires d'images. Les 6 paramètres des équations 2.2 et 2.3 sont alors estimés par optimisation. Dans notre cas nous avons utilisé la technique d'optimisation non linéaire "Levenberg-Marquardt" [47].

Au final, ces paramètres nous permettent de générer une correspondance pixel à pixel de l'image de profondeur vers l'image thermique. Cette relation est non-réciproque du fait de l'information du pixel qui est différente pour chaque image.

2. 4 Bases de données et environnement d'acquisition

Les techniques basées sur l'apprentissage sont très gourmandes en bases de données. Et même dans le cas de techniques de traitement d'images plus classiques, des bases de données sont nécessaires pour l'évaluation des performances pour la tâche assignée. Dans notre cas plus particulier, le suivi d'une personne âgée et l'estimation de son activité à partir d'images de profondeur et d'images thermiques basses résolutions, les bases de données publiques sont relativement rares. Le plus souvent ces bases de données sont focalisées sur la détection de chute avec ceci de particulier que du fait de la population surveillée et pour des raisons d'éthique et de respect de la vie privée, il existe très peu de

bases de données publiques sur des chutes réelles. Ainsi, la majorité des algorithmes a été évaluée sur des chutes simulées par des personnes plus jeunes.

Dans la suite de cette section, nous allons, dans un premier temps, faire l'inventaire des quelques bases de données publiques axées sur la détection de chutes à partir d'images de profondeur ou d'images thermiques. Dans un second temps, nous allons présenter les bases de données réalisées dans le cadre de notre projet.

2. 4.1 Bases de données publiques

2. 4.1.1 Base de données d'images de profondeur

2. 4.1.1.1 UR Fall Detection Dataset. En 2014, Kwolek et Kepski [53] ont crée une base de données pour l'estimation de chutes à partir d'images de profondeur. Des scènes de vie quotidienne se déroulant dans une pièce et des chutes simulées ont été acquises à l'aide de deux caméras Kinect. La première caméra a été placée en face de la personne à une distance d'un mètre du sol et la deuxième caméra a été fixée au plafond à une distance de 2,5 mètres du sol pour couvrir toute la pièce. Cette base contient 70 séquences : 30 séquences simulant des chutes et 40 séquences représentent les activités humaines de la vie quotidienne (Activities of Daily Living ADL) qui ont été enregistrées par une seule caméra. Ces simulations ont été réalisées par cinq personnes en bonne santé en essayant deux types de chutes : chute en position debout et chute en position assise sur une chaise. La Figure 2.12 illustre quelques exemples de cette base de données.

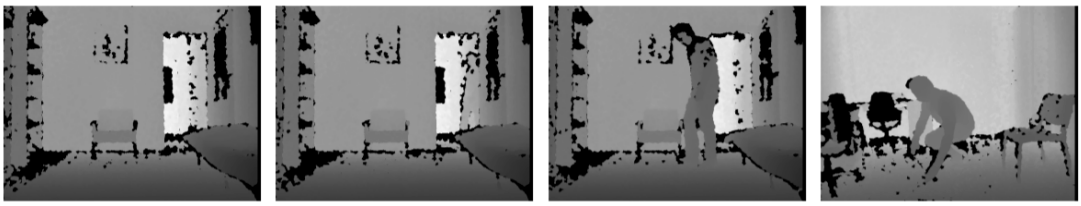


FIGURE 2.12 – Exemples de la base "UR Fall Detection" [53].

Cette base de données est accessible à l'adresse suivante :
http ://fenix.univ.rzeszow.pl/mkepski/ds/uf.html

2. 4.1.1.2 Base de données dans des conditions réelles en EHPAD. Stone et al. [54] ont installé un dispositif, constitué d'une caméra kinect et d'un ordinateur portable,

dans 16 studios d'un Ehpad où les résidents sont âgés de 67 à 97 ans (7 hommes et 9 femmes). La Figure 2.13 montre l'installation de la caméra Kinect dans un appartement. La caméra Kinect est placée sur une petite étagère à quelques centimètres du plafond (à une hauteur de 2,75 m du sol), au-dessus de la porte d'entrée. L'ordinateur est placé dans une armoire au-dessus du réfrigérateur. Selon les auteurs, le choix du placement des équipements est très important car ils doivent rester discrets vis-à-vis des résidents.

FIGURE 2.13 – Installation du dispositif de détection de la chute dans l'appartement d'une personne âgée [54].

Pour traiter l'information de profondeur, les chercheurs ont choisi d'utiliser les valeurs brutes de profondeur obtenues via la bibliothèque open source libfreenect [55], plutôt que des coordonnées du squelette de la personne dans la scène fournis par le kit logiciel SDK de Microsoft. Quelques exemples d'images de profondeur sont présentées dans la Figure 2.14. Malheureusement, cette base de données n'est pas accessible.

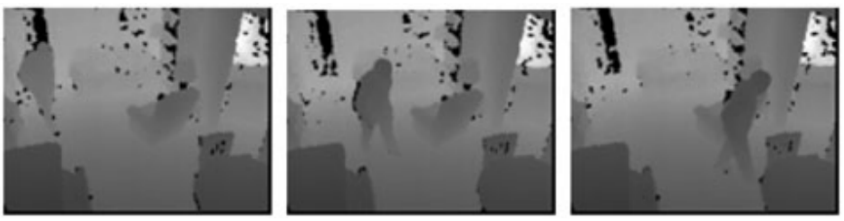


FIGURE 2.14 – Exemples de la base de données réelle [54].

2. 4.1.1.3 SDU Fall Dataset. Cette base a été réalisée à l'aide de dix jeunes hommes et femmes qui ont répété six types d'activités 30 fois, ce qui a donné 1800 séquences. Parmi ces enregistrements, 1197 vidéos ont été validées, dont 997 séquences d'activité de la vie quotidienne et 200 chutes. Ces vidéos ont été capturées à l'aide d'une caméra Kinect avec une fréquence de 30 FPS et une résolution de 320 × 240 pixels, stockées au format AVI [56]. La Figure 2.15 présente quelques images de cette base de données.

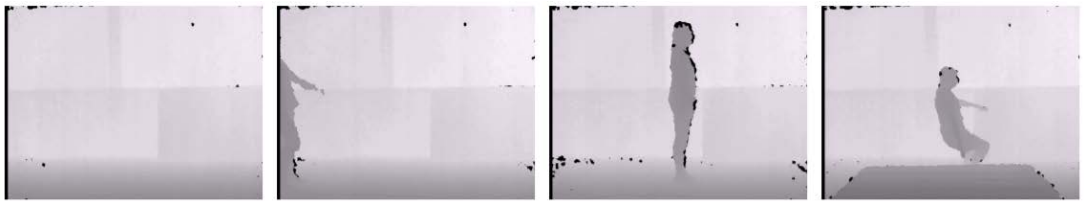


FIGURE 2.15 – Exemples de la base de donnés "SDU Fall Dataset" [56].

2. 4.1.1.4 Base de données de Mastorakis et Makris. Mastorakis et Makris [57] ont simulé des chutes à l'aide d'une Kinect fixée sur un trépied à 2,04 m du sol et à une distance maximale de 7 m de la zone de chute possible. Ils ont réalisé 184 vidéos réparties comme suit :

— 48 séquences relatives à des chutes,
— 12 séquences liées à des chutes lentes,
— 48 séquences correspondant à une position allongée,
— 32 séquences pour la position assise,
— 12 activités avec une vitesse lente,
— 32 séquences pour lesquelles la personne est en mouvement de ramassage des objets par terre.

Les séquences contiennent des chutes : vers l'avant, vers l'arrière et de côté. Huit personnes ont participé à la création ces données, deux d'entre elles ont fait toutes les activités avec un rythme ralenti pour simuler le comportement d'une personne âgée. Cette base n'est pas publique.

2. 4.1.1.5 TST Fall Detection dataset. L'ensemble des données est composé de séquences d'activité de la vie quotidienne (ADL) et des chutes simulées par 11 jeunes acteurs. Ces séquences contiennent des images de profondeur de taille 512×424, capturées par la caméra Kinect V2.

Les activités suivantes font partie de la catégorie ADL :

— la personne est assise sur une chaise ;
— la personne marche puis elle ramasse un objet par terre ;
— la personne fait des aller - retours ;
— la personne s'allonge.

Les activités suivantes font partie de la chute :

— **En avant :** l'acteur s'écroule par terre suite à un glissement ou à un heurtement d'obstacle.
— **En arrière :** l'acteur tombe en arrière surtout pour simuler les personnes qui ont une maladie de Parkinson ou qui présentent un "état lacunaire" (trouble de la marche).
— **De côté :** l'acteur tombe sur le côté pour représenter les chutes par malaise.
— **EUpSit :** l'acteur tombe en arrière et il finit par s'asseoir.

Chaque acteur a répété chaque action 3 fois, générant un nombre total de 264 séquences, mais elles aussi ne sont pas publiques.

2. 4.1.2 Base de données thermiques

2. 4.1.2.1 Thermal Simulated Fall. Pour leur étude réalisée en 2017 sur la détection de chute par un capteur thermique, Vadivelu et al ont réalisé des séquences d'activités dans une petite pièce. Cette base de données est inspirée de la base "KUL SIMUALTED FALL DATASET" (cette dernière est un benchmark fait avec des images couleurs). Elle contient 9 séquences de la vie quotidienne (ADL) et 35 séquences de chute. Ces séquences sont enregistrées à l'aide de la caméra thermique FLIR ONE de haute résolution (640×480) montée sur un smartphone de type Android. Les auteurs ont ainsi pu démontrer que le capteur thermique présentait un avantage majeur sur le plan du respect de la vie privée et de l'intimité en protégeant l'identité de l'individu et en le capturant même en l'absence de lumière. La Figure 2.16 présente un échantillon d'images. Cette base de données est mise à la disposition du public [27].

2. 4.1.2.2 eHomeSeniors. Cette base de données est publique [58]. Elle contient un total de 445 scénarios de 15 chutes différentes simulés par un groupe de six personnes. Ce groupe contient trois acteurs qui ont suivi les instructions d'un physiothérapeute pour simuler de vraies chutes. Cette base est enregistrée à l'aide de deux types de capteurs thermiques, Figure 2.18 :

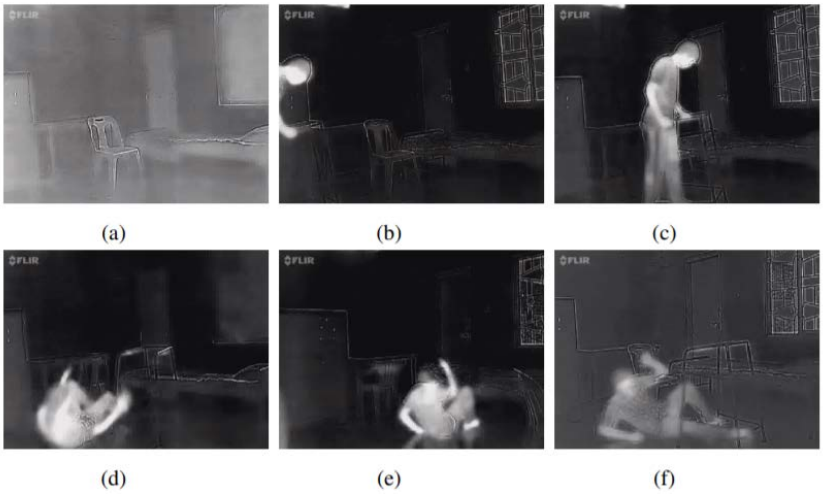


FIGURE 2.16 – Exemple d'images fournies par le capteur thermique Flir One (a : scène vide , b : la personne entre dans la scène, c : exemple d'ADL et d,e,f : exemples des chutes) [58].

— **Capteur Melexis MLX90640.** Il est peu coûteux (environ 50 euros) et il fournit une image de basse résolution (32×24 pixels), avec une fréquence d'environ 16 fps. Il est capable de mesurer la température des objets entre -40°C et 300°C. La Figure 2.17 montre deux exemples d'images acquises par ce capteur. Ce capteur a été fixé à une hauteur de 1,2 m, de sorte que l'angle de vision est réparti équitablement à partir du centre du capteur, formant un angle vertical de 75° et un angle horizontal de 110° , comme le montre la Figure 2.17.

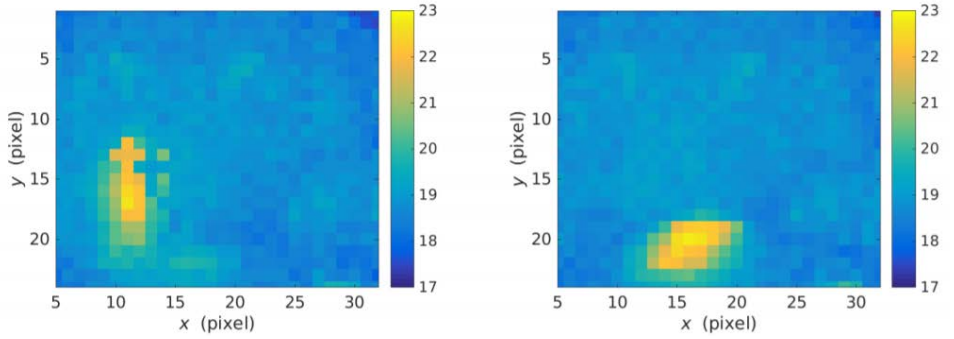


FIGURE 2.17 – Exemple d'images fournies par le capteur Melexis (rouge : la température la plus élevée, bleu : la température la moins élevée) [58].

— **Capteur Omron D6T-8L-06.** Il fournit une image de très basse résolution (1×8 pixels) qui peut mesurer la température des objets entre -10°C et 60°C. Afin d'identifier les chutes, quatre capteurs de ce type ont été utilisés : deux capteurs

à mi-hauteur (1 m du sol) et deux autres au niveau du sol (10 cm du sol). La chute est reconnue à la fois par la diminution de la température identifiée par les capteurs supérieurs et par l'augmentation de la température par les capteurs inférieurs. L'emplacement de ces capteurs par rapport au premier est illustré par la Figure 2.18.

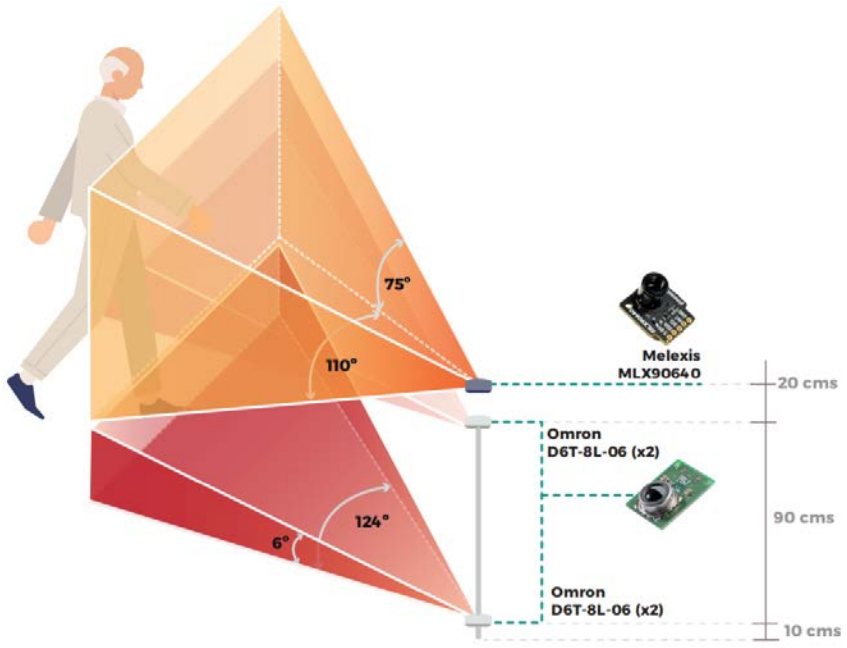


FIGURE 2.18 – Position des capteurs dans la scène [58].

2. 4.1.3 Synthèse des bases de données publiques

Pour résumer, il existe plusieurs travaux basés sur les données de profondeur (des kinects dans tous les cas) et d'autres travaux (beaucoup moins) basés sur les données thermiques relatives à la détection de chutes des personnes âgées. La Table 2.4 récapitule toutes les bases de données mentionnées précédemment. L'utilisation de ces bases de données nous pose de nombreux problèmes. Les données ne sont généralement pas disponibles en archives ouvertes. Le nombre de séquences est généralement réduit, ce qui ne permet pas des approches de type apprentissage. De surcroît, si les données de profondeurs sont issues du même capteur que celui que nous avons envisagé, cela n'est pas le cas pour le capteur thermique. Aucune de ces bases de données ne permet la fusion de l'information de nos deux capteurs. Pour cette raison, nous avons décidé de créer notre base de données.

Base de données	UR Fall Detection	Base de données en Ehpad	SDU Fall Dataset	Base de Mastorakis et Makris	TST Fall Detection	Thermal Simulated Fall	eHome Seniors
Disponible en accès libre	Oui	Non	Non	Non	Non	Oui	Non
Type de capteurs	2 Kinect	1 Kinect	1 Kinect	1 Kinect	1 Kinect V2	1 Flir One	1 Melexis et 4 Omron
Séquences	70 (40 ADL et 30 chutes)	Vidéos de 3339 jours	1197 (997 ADL et 200 Chutes)	184 (124 ADL et 60 chutes)	264 (132 ADL et 165 chutes)	44 (9 ADL et 35 chutes)	445 chutes
Personnes	5 personnes en bonne santé	16 personnes âgées (7 hommes et 9 femmes)	10 jeunes	8 personnes en bonne santé	11 personnes en bonne santé	NA	6 personnes en bonne santé
Type de chutes	Deux chutes (à partir de la position debout et assise)	Plusieurs types de chutes	Six types de chutes	Chute lente et chute normale	5 types de chutes	NA	15 types de chutes
Environnement	Non réel	Réel	Non réel	Non réel	Non réel	Non réel	Non réel
Année	2014	2014	2014	2014	2014	2017	2019

TABLE 2.4 – Bases de données pour la détection de chutes des personnes âgées.

De plus, comme il a été dit précédemment, les bases de données publiques ne permettent pas d'évaluer notre étude, étant donné qu'il n'existe aucune base publique contenant l'image thermique et de profondeur à la fois avec les paramètres de calibration. Pour cette raison, nous avons décidé de créer notre base de données.

2. 4.2 Bases de données créées pour le projet

2. 4.2.1 Scénario

Dans le projet PRuDENCE, la détection précoce de la dégradation de la locomotion d'un individu est définie comme un autre objectif pour prédire des changements de mode de vie de la personne âgée. Cette prévention devra s'appuyer sur une analyse de

caractéristiques dynamiques telles que l'équilibre, la posture, l'activité ou la démarche de l'individu. Cela aussi nécessite une base de données propre à notre contexte, ce qui nous a conduit à l'acquisition de nos propres séquences en se basant sur les scénarios proposés par le protocole d'évaluation des performances de détecteur de chute du Technopôle Alpes, Santé à Domicile et Autonomie (TASDA) [59]. Ces scénarios ont été créés pour les capteurs portables. Alors, nous avons modifié la liste d'activités, selon nos besoins, pour avoir 37 scénarios (20 représentant les chute et 17 l'activité normale). La Table 2.5 définit les différents types d'activités. Nous avons fait l'acquisition de ses séquences en collaboration avec l'Université de Technologies de Troyes (UTT) dans un Living Lab ActivAgeing (LL2A), voir Figure 2.19, en présence d'un kinésithérapeute, des personnes âgées et des acteurs.

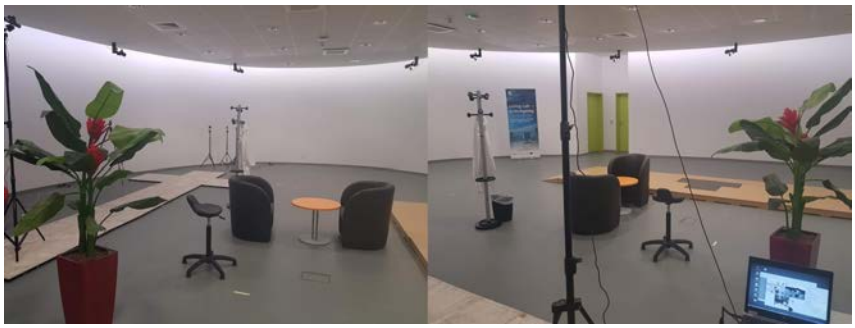


FIGURE 2.19 – Living Lab ActivAgeing (LL2A) à l'UTT.

Le LL2A est spécialisé en co-conception et évaluation par les usages des solutions pour l'autonomie des personnes âgées. Il possède sa plateforme de recherche et d'expérimentation au sein de l'Institut Charles Delaunay (UMR CNRS 6281) et un logement laboratoire-démonstrateur. La plateforme présente la spécificité d'intégrer un dispositif d'analyse d'activité par codage vidéo temps réel avec une centrale d'acquisition de données physiologiques ambulatoire couplée d'un système d'analyse tridimensionnelle du mouvement, appelé Vicon Motion Capture System.

2. 4.2.2 Bases de données

Nous avons fait les acquisitions dans trois locaux différents : le Living lab LL2A, un logement démonstrateur de l'UTT (voir Figure 2.20) et un logement à l'ECAM (voir Figure 2.21).

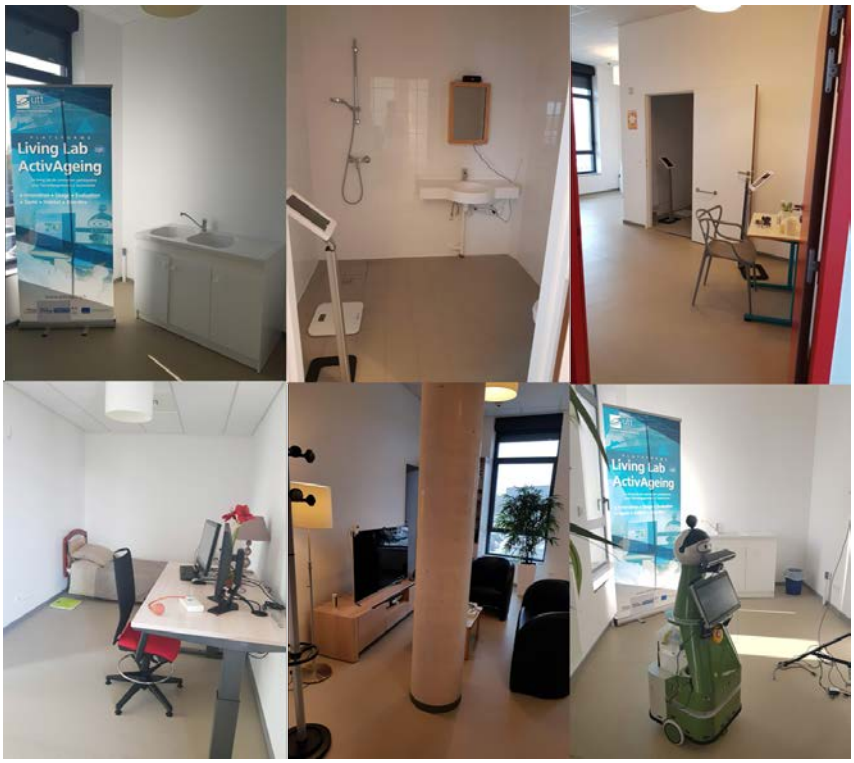


FIGURE 2.20 – Appartement d'acquisition à l'UTT.

Au LL2A, nous avons enregistré les situations décrites dans la Table 2.5, avec une fréquence de 8 Hz, en répétant chaque situation deux fois par deux acteurs différents (un kinésithérapeute et un jeune). Les acteurs sont visibles par les deux systèmes d'acquisition (Vicon et notre dispositif). Ils ont été identifiés par le système Vicon à partir des marqueurs lumino-réfléchissants posés sur la peau (voir Figure 2.22). L'utilité de ce système est l'annotation automatique de la base de données.

FIGURE 2.21 – Appartement d'acquisition à l'ECAM.

Dans les deux autres locaux, quatre jeunes ont fait les enregistrements de 37 scénarii d'une durée de 60 secondes chacun et avec une fréquence de 2 Hz. Pour simuler une journée d'une personne âgée, six acquisitions ralenti ont été faites dans l'appartement de l'UTT par trois jeunes d'une durée de 30 minutes chacune et avec une fréquence de 2 Hz. environnement plus réaliste, scènes de la réalité.

Dans ces conditions, nous avons enregistré 139 séquences de différentes tailles. La Table 2.6 résume la répartition de ces séquences enregistrées dans le cadre de cette thèse. Notre base est publique (https ://drive.google.com/drive/folders/1-hWFHicZLdoYMUnGkN00I762582R4-1P ?usp=sharing).

FIGURE 2.22 – Des marqueurs lumino-réfléchissants posés sur un kinésithérapeute.

2. 5 Conclusion

Dans ce chapitre, nous avons présenté, dans un premier temps, le contexte général de cette thèse, ensuite nous avons justifié le choix des capteurs. Dans un second temps, nous avons détaillé l'étape de calibration qui permet de combiner les deux informations issues de deux types de capteurs ainsi que la création de notre propre mire. Enfin, nous avons étudié les bases de données publiques pour finir par créer notre propre base.

Situations de chutes classiques/rapides (à forte accélération)	
1	Chute en avant sur les genoux
2	Chute en avant se terminant en position allongée à plat
3	Chute en avant avec rotation gauche se terminant en position allongée latérale gauche
4	Chute en avant avec rotation droite se terminant en position allongée latérale droite
5	Chute en arrière se terminant en position assise
6	Chute en arrière se terminant en position allongée à plat
7	Chute en arrière avec rotation gauche se terminant en position allongée latérale gauche
8	Chute en arrière avec rotation droite se terminant en position allongée latérale droite
9	Chute latérale à gauche
10	Chute latérale à droite
Situations de chutes lentes/molles (à faible accélération)	
11	Chute molle en retenant à un lit
12	Chute molle en retenant à une table
Situations de la vie quotidienne (ADL)	
13	S'assoir sur une chaise, marquer une pause et se lever
14	S'allonger sur un lit , marquer une pause et se lever
15	Marcher dans un appartement
16	Se baisser, ramasser un objet au sol et se relever
17	Boire un verre d'eau ou son café
18	Lire le journal
19	Mesurer la tension
20	Regarder la télé
21	Marcher dans le couloir, ouvrir la porte pour recevoir son invité et s'installer dans le salon
22	Laver la vaisselle
23	Préparer à manger
24	S'habiller pour sortir

TABLE 2.5 – Situations typiques des activités faites par la personne âgée chez elle.

Lieu	Nombre de séquences	Acteurs	Durée de chaque séquence	Fréquence
Living Lab LL2A	96	2	de 5 à 10 s	8Hz
Appartement de l'UTT	6	3	30 minutes	2Hz
Appartement de l'ECAM	37	4	60 s	2Hz

TABLE 2.6 – Séquences enregistrées dans le cadre de cette thèse.

Deuxième partie

Suivi de la personne avec des capteurs basés sur la vision

Chapitre 3

État de l'art

Dans le premier chapitre, nous avons étudié les différents types de chutes des personnes âgées, leurs causes et leur impact sur la société ainsi que leurs nombreux systèmes de détection. Dans le deuxième chapitre, nous avons justifié notre choix de capteurs de profondeur et thermique en fonction de l'étude faite dans le chapitre précédent. Le troisième chapitre de cette thèse est consacré à détailler la manière dont nous avons traité les images issues de ces capteurs pour fournir des indicateurs de dégradation physique de la personne âgée dans un contexte de détection et de la prévention des chutes. En effet, ces indicateurs permettent de détecter les chutes chez les personnes âgées à domicile. Dans cette section, nous détaillons tout d'abord les différentes analyses faites sur les images pour extraire la personne dans la scène. Ensuite, nous citons les méthodes de suivi d'une personne. Enfin, nous décrivons le principe de filtrage bayésien, plus précisément le filtre particulaire qui est adapté à notre solution.

3. 1 Modèle de suivi d'un objet en mouvement

Dans notre contexte, l'objectif du projet PRuDENCE est d'assurer la sécurité des personnes âgées indépendantes à domicile. Pour cette raison, nous nous intéressons plus particulièrement au suivi d'une seule personne en mouvement. En effet, une personne accompagnée est considérée en sécurité. Dans la suite de ce chapitre, nous étudions les principaux modèles du suivi d'un objet en mouvement qui existent dans la littérature.

La Figure 6.11 décrit le processus général d'un modèle de suivi. Le choix des capteurs, la calibration et l'acquisition des images, détaillés dans le chapitre précédent, représentent le pré-traitement nécessaire pour fournir une information utile. Dans cette section, nous nous intéressons à la deuxième partie *analyse de données*. Celle-ci implique la détection et le suivi d'objets. L'étape de la détection d'objet en mouvement est importante pour localiser la silhouette de la personne dans la scène. Elle consiste à identifier les pixels représentant l'objet et à regrouper ces pixels sous forme d'une région d'intérêt (Region

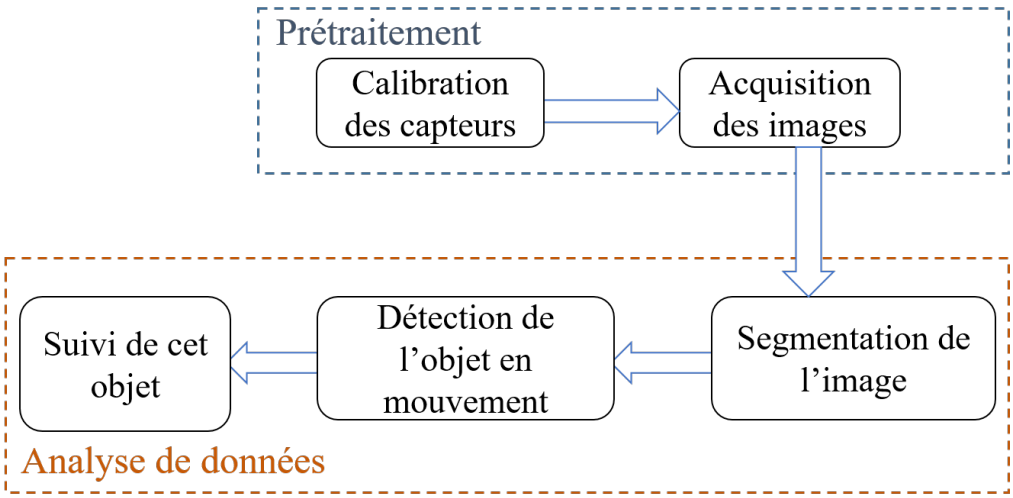


FIGURE 3.1 – Processus de suivi d'un objet en mouvement.

Of Interest ROI). La dernière étape de ce processus est consacrée à la sélection de la région d'intérêt et au suivi du mouvement. Dans la suite, nous définissons les méthodes appliquées dans chaque étape en prenant en compte le type des capteurs utilisés dans notre contexte.

3. 1.1 Techniques de détection d'objet

La détection de la personne dans la scène s'avère l'étape primordiale pour notre étude. Le principe de cette étape consiste à éliminer le bruit et ne garder que les pixels qui contiennent une information utile appelée région d'intérêt. Ces régions peuvent être extraites à partir de la technique de détection des contours ou de la segmentation de l'avant plan. En outre, l'étape de segmentation est très intéressante si son temps d'exécution ne ralentit pas le traitement du système. Nous pouvons classer les méthodes de détection des régions d'intérêts en fonction de la manière de traiter l'information. Deux façons sont envisageables (Figure 3.2) : (i) Extraire des points susceptible d'être intéressants ou (ii) analyser le mouvement des pixels mobiles.
La première catégorie consiste à extraire des points d'intérêt en local (en analysant des régions de l'image) ou global (en analysant toute l'image). La deuxième catégorie est dédiée plutôt à détecter et analyser le mouvement des pixels en mouvement. Puisque la chute est caractérisée par un mouvement rapide, il sera intéressant pour notre étude d'adopter la deuxième catégorie.

Dans cette section, nous définissons brièvement quelques méthodes d'extraction des points d'intérêt. Cependant, nous détaillons les méthodes qui se concentrent sur le mouvement de l'objet. Ce choix est justifié par le fait que ces méthodes sont de plus en plus utilisées dans le contexte de la détection d'un objet en mouvement sur des images de profondeur.

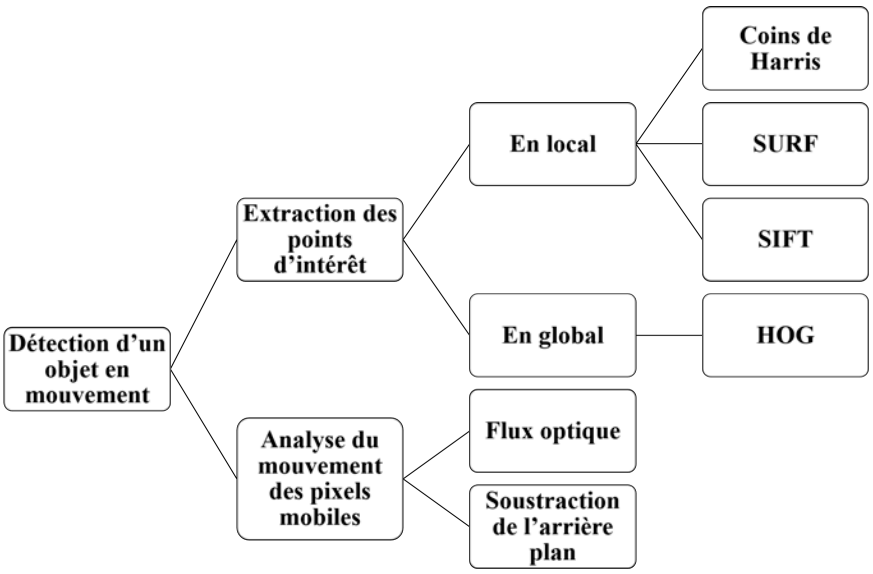


FIGURE 3.2 – Méthodes utilisées pour la détection d'un objet en mouvement.

3. 1.1.1 Extraction des points d'intérêt

Dans la première catégorie, les méthodes sont basées sur l'extraction des points d'intérêt localement, sur une simple portion de l'image, ou globalement en traitant toute l'image. La technique de traitement local consiste à analyser l'état des pixels d'intérêt ainsi que leurs voisinages. Dans ce contexte, le terme descripteur de caractéristiques est souvent utilisé, désignant la structure de données qui est utilisée afin de calculer la similarité entre deux points caractéristiques. Différentes méthodes ont été proposées dans la littérateur dont les trois suivantes sont les citées :

— **Le détecteur de coin de Harris (Harris corner) :** c'est l'opérateur de points d'intérêt le plus populaire. Les coins sont des régions de l'image qui présentent de grandes variations d'intensité dans toutes les directions. Une première recherche pour trouver ces coins a été faite par Chris Harris et *al.* [60]. Le détecteur fonctionne

en calculant des dérivées horizontales et verticales de l'image et en recherchant des zones où les deux dérivées sont élevées.

— **SIFT (Scale Invariant Feature Transform) :** Le principe de ce descripteur est basé sur le calcul d'un histogramme de gradients locaux orientés autour d'un point. Les caractéristiques sont invariantes à l'échelle et à la rotation de l'image, et partiellement invariantes aux changements d'éclairage et de point de vue de la caméra. Toutefois, ce discripteur est coûteux en temps de calcul [61].

— **SURF (Speeded Up Robust Features) :** c'est un descripteur plus rapide et robuste par rapport au descripteur SIFT. Il fonctionne en appliquant un masque approximant la dérivée gaussienne sur l'image à plusieurs échelles. La méthode est très rapide grâce à l'utilisation du masque où la valeur d'un pixel (x,y) est la somme de toutes les valeurs dans le rectangle défini par l'origine de l'image (0,0) et les paramètres (x,y) [62].

L'analyse globale traite l'état de tous les pixels de l'image. La méthode la plus connue est :

— **l'histogramme de gradients orientés (HOG) :** Ce descripteur se concentre sur la forme de l'objet. Il extrait le gradient et l'orientation des pixels du conteur de l'objet [63]. En outre, ces orientations sont calculées en portions "localisées". Cela signifie que l'image est décomposée en petites régions et les valeurs de gradient et de l'orientation sont calculées pour chaque région. Les histogrammes sont créés en utilisant ces valeurs, ce qui justifie son nom "Histogramme des gradients orientés". Par exemple, Yucai et. al [64] ont utilisé le descripteur HOG pour un robot de service afin de détecter et suivre des personnes à l'intérieur.

3. 1.1.2 Analyse du mouvement de l'objet

Dans la deuxième catégorie, les méthodes, telles que le flux optique ou l'extraction du fond, traitent le déplacement de chaque pixel durant au moins deux images consécutives pour détecter les zones en mouvement.

3. 1.1.2.1 Flux optique : le flux optique est souvent utilisé dans les algorithmes de détection d'un objet en mouvement. L'idée de cette méthode est d'estimer le mouvement de la scène ou de l'objet en se basant sur le calcul de vecteurs de déplacement de chaque pixel durant deux images successives. Les méthodes différentielles estiment le flux optique sous l'hypothèse d'une luminosité constante, ce que n'est pas le cas pour notre projet.

Si l'intensité d'un pixel à la position (x, y) et à l'instant t est désignée par $I(x, y, t)$, la contrainte de luminosité est désignée par l'équation 3.1 :

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{3.1}$$

où $(\delta x,\ \delta y)$ correspond au déplacement du point (x, y) dans l'intervalle de temps δt. En supposant que les variations sont très faibles, un développement de Taylor du premier ordre de l'équation 3.1 donne :

$$\frac{\partial I}{\partial x}d_x + \frac{\partial I}{\partial y}d_y + \frac{\partial I}{\partial t}d_t = 0 \tag{3.2}$$

$$\frac{\partial I}{\partial x}d_x + \frac{\partial I}{\partial y}d_y = -\frac{\partial I}{\partial t}d_t \tag{3.3}$$

En notant $\frac{d_x}{d_t} = \dot{x}$ et $\frac{d_y}{d_t} = \dot{y}$ la vitesse du point dans l'image, $\nabla I_x = \frac{\partial I}{\partial x}$ et $\nabla I_y = \frac{\partial I}{\partial y}$ le gradient spatial de l'intensité de ce point et $\frac{\partial I}{\partial t} = \dot{I}$ la variation temporelle de la luminosité, l'équation 3.3 s'écrit comme suit :

$$\nabla I^x.\dot{x} + \nabla I^y.\dot{y} = -\dot{I} \tag{3.4}$$

L'équation (3.4) est appelée l'équation de contrainte de mouvement en 2-D ou de contrainte de gradients. Cela mène à un système linéaire à deux inconnues ($\dot{x}$ *et* $\dot{y}$) qui doivent être estimés à partir d'une seule équation[65]. Pour résoudre ce problème, une condition supplémentaire doit être introduite. Horn et Schunck [66] ont utilisé une condition de lissage global. Ils ont appliqué une cohérence spatiale légère en forçant les dérivées partielles des vecteurs de mouvement voisins à être minimales. Une cohérence spatiale intense a été introduite par Lucas et Kanade [67]. Celle-ci estime que le mouvement dans une petite région est constant.

Le flux optique s'applique uniquement sur la zone qui contient l'objet à analyser. Gao et al. [68] utilisent par exemple le flux optique pour suivre les avant-bras et la tête d'une personne âgée prenant son repas. Dans les travaux de Jean et al. [69], le flux optique est employé dans le but de suivre uniquement les pieds d'une personne. Cependant, Senst et al. [70] ont montré que la méthode de flux optique reste une méthode complexe et très

coûteuse en temps de calcul quelle que soit l'implémentation choisie, ce qui n'est pas le choix le plus pertinent pour les systèmes de détection des chutes. De plus Hu et al. [71], Li et Yu [72] et Jemilda et al. [73] ont remarqué que cette méthode est sensible au bruit et elle nécessite du matériel robuste pour le traitement en temps réel.

3. 1.1.2.2 Soustraction de l'arrière-plan De nombreuses applications [74] se servent d'une caméra statique pour détecter et suivre des personnes ou des objets en mouvement dans des séquences vidéo. Ces applications utilisent généralement des algorithmes de soustraction de l'arrière-plan. Cet algorithme sépare l'avant-plan, qui contient les objets mobiles, de l'arrière-plan. La technique consiste à apprendre un modèle d'arrière-plan et à classer tous les objets qui n'appartiennent pas à ce modèle comme avant plan. La soustraction de l'arrière-plan est basée sur la modélisation du fond illustré par une carte de référence. Cette dernière est comparée à chaque image de la séquence afin de déterminer les variations possibles. Une mise à jour de la carte peut s'effectuer pour actualiser l'arrière plan. Cette étape de pré-traitement permet de réduire la complexité des analyses ultérieures et peut même améliorer le résultat final. Généralement, un modèle d'arrière-plan est créé sur la base d'une scène vide qui ne contient aucun objet en mouvement. Cannons [75] a évalué des nombreux algorithmes de soustraction d'arrière-plan basés sur des images couleur. Les modèles les plus utilisés sont le modèle Gaussien SGM (Single Gaussian Model en anglais) et le mélange de Gaussiennes MOG (Mixture Of Gaussian en anglais).

— **Un modèle gaussien (SGM) :** Dans cette méthode, la scène est modélisée par l'hypothèse que l'historique des valeurs d'intensité d'un pixel peut être modélisé par une gaussienne (Single Gaussian Model SGM). Á l'étape d'apprentissage, les valeurs moyennes et de variance de chaque pixel sont calculées. Ensuite, les pixels ayant des moyennes supérieures à l'écart type sont considérées comme étant l'avant plan. Cette méthode a été utilisée dans Pfinder [76]. Wren et al. ont utilisé le modèle SGM pour modéliser l'arrière plan.

— **Le mélange de gaussiennes (MOG) :** Ce modèle est dédié plutôt aux arrière-plans complexes. Il peut être facilement adapté aux changements brusques dans le fond, comme l'éclairage au niveau des images couleurs. Hayman et al. [77] ont utilisé un mélange de Gaussien (MoG) qui est plus approprié pour représenter un arrière-plan complexe. Zivkovic et al. [78] ont proposé un MoG adaptatif, qui est capable de modifier les paramètres du modèle pour de nouvelles scènes.

De nombreux cas pratiques, basés sur la méthode de détection du fond, sont utilisés notamment pour les systèmes de détection et de suivi des personnes [79, 80, 81] et même pour la détection des chutes chez les personnes âgées [82, 33, 83]. Par exemple, Schwarz et al. [49] ont simplement utilisé la soustraction statique de la première image. Par ailleurs, Jansen et al. [84] ont créé un modèle d'arrière-plan comme une moyenne des premières images vides. Quant aux travaux de Guomundsson et al. [85], ceux-ci ont caractérisé l'arrière-plan avec le modèle gaussien probabiliste. Ce modèle a été appliqué par Harville et al. [86] qui adoptent la méthode de segmentation gaussienne pour extraire l'avant plan des images RGB-D.

3. 1.2 Suivi d'un objet en mouvement

Le suivi d'un objet en mouvement dans une séquence d'images concerne une estimation des positions de cet objet. Le traitement d'un tel sujet complexe et très vaste est difficile. Par exemple, l'environnement d'acquisition, la forme de l'objet, qui peut varier notamment s'il s'agit de personnes, et le bruit représentent des complications qu'il faut traiter pour avoir une méthode de suivi robuste. Certaines méthodes de suivi sont très spécifiques et destinées à être utilisées dans des environnements bien contrôlés et à faible taux de bruit. D'autres algorithmes de suivi sont plus généralisés, précis et plus robustes à des occultations ou des interférences mais ils exigent un effort de calcul plus important.

Dans le contexte de la détection des chutes chez les personnes âgées, le comportement de la personne peut être un élément intéressant pour détecter les situations à risques y compris les chutes en premier temps et contrôler ces situations dans un second temps. Le suivi permet de fournir des indicateurs utiles pour surveiller le comportement de la personne. Certains travaux ont été réalisés pour détecter la personne sur des images RGB. D'autres préfèrent utiliser les images de profondeur pour respecter la vie privée de l'individu en surveillant le comportement du corps humain. D'autres encore décident d'utiliser les capteurs thermiques puisqu'ils sont moins chers que ceux de profondeur.

Le choix de la région d'intérêt dépend principalement de l'événement à détecter. Nous allons suivre, dans notre cas, la personne afin d'extraire des paramètres fiables pour analyser et interpréter le comportement de cette personne dans une scène quelconque. Pour cette raison, le suivi de la personne peut se faire en ne suivant qu'une seule partie de sa silhouette.

Dans ce paraphe, nous allons étudier en détail les techniques de suivi existantes. Nous avons classé les méthodes de suivi de la personne en trois catégories en fonction de la

représentation de la personne (soit le centre de la tête, ou la silhouette ou bien le centre de la masse) extraite sur des images de vision (images couleurs, de profondeur et thermique) : les méthodes basées sur 1) l'analyse du mouvement de la tête, 2) l'analyse de la silhouette et 3) le suivi de l'inactivité (Figure 3.3).

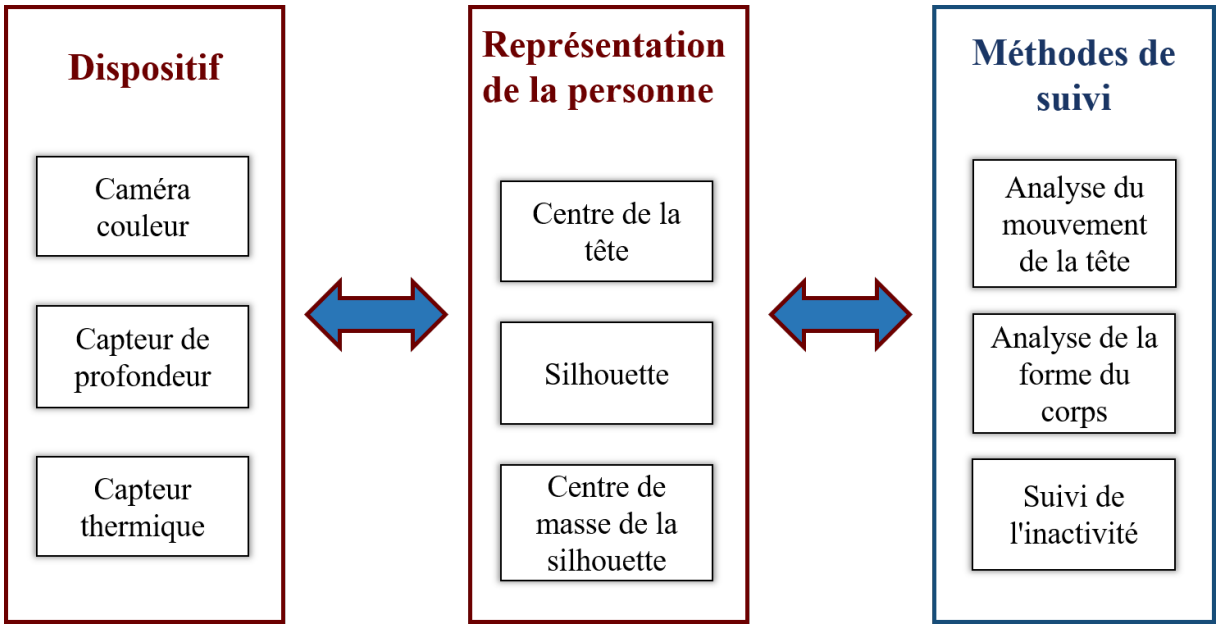


FIGURE 3.3 – Méthodes de suivi basées sur les systèmes de vision.

3. 1.2.1 Analyse du mouvement de la tête

La représentation de la personne par le centre de la tête a été utilisée dans plusieures applications en raison de son mouvement vertical qui est important lors d'une chute. De plus, la tête est la partie la plus haute du corps. Elle est souvent visible dans la scène. L'analyse des mouvements de la tête peut donc fournir un indice important pour la détection des chutes. Par exemple, Rougier et *al.* [87] ont développé un détecteur de chute basé sur le suivi de la tête sur des images de profondeur. Un paramètre crucial de cette méthode est la vitesse de la tête qui est souvent importante lors d'une chute. Dans leur article [88], Nghiem et *al.* ont proposé un algorithme de suivi de la tête basé sur des images de profondeur. L'algorithme estime les positions possibles de la tête, puis il détecte la position finale en repérant la position des épaules. La recherche des positions est rapide puisque le traitement ne se fait que sur la partie supérieure de la silhouette de la personne. La reconnaissance des personnes est basée sur l'histogramme de gradient

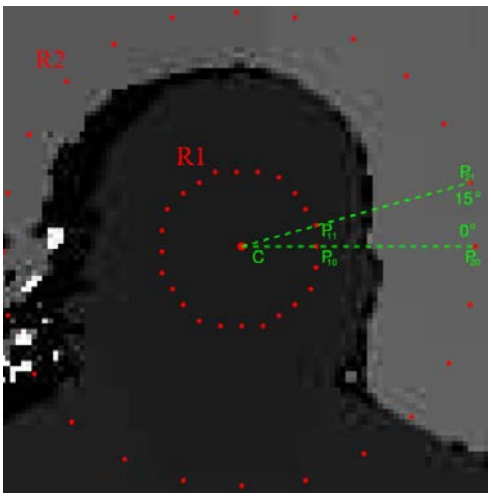


FIGURE 3.4 – Descripteur circulaire pour détecer la tête d'une personne [89].

de la tête et des épaules. Cet algorithme est plus robuste à l'articulation humaine et à la flexion du dos que le HOG original. En outre, le HOG modifié est dépendant de la vitesse de la tête ainsi que de la distance du centre de la tête par rapport au sol. En utilisant à la fois les centres de masse de la silhouette et de la tête, cet algorithme est moins affecté par les variations du centroïde. Chan et al. [89] ont proposé un descripteur de tête circulaire, comme le montre la Figure 3.4, pour classer chaque pixel de l'image appartenant à la tête. Si un pixel C est proche du centre de la tête, une fenêtre circulaire (R1) se trouve entièrement à l'intérieur de la tête et l'autre (R2) entoure complètement la tête. Une haute performance de détection des têtes (90% de précision) a été atteinte pour une base de données faite sur mesure, notamment pour des personnes proches et en face de la caméra. Par contre, pour un ensemble de données publiques composé de quelques activités quotidiennes, la précision baisse et vaut environ 70%.

Récemment, Ballotta et al. ont exploité une nouvelle méthode de détection de la tête sur des images de profondeur qui se base sur une approche d'apprentissage profond [90]. Cette approche trouve ses limites puisqu'elle ne traite pas les séquences en temps réel. Pour cette raison, Ballotta et al. ont développé d'autres études afin d'améliorer cette version [91]. En particulier, le système présenté remplace l'approche classique des fenêtres coulissantes par un réseau entièrement convolutif. Deux ensembles de données publiques sont utilisés pour apprendre et tester le réseau proposé. Les résultats expérimentaux confirment l'efficacité de la méthode comparée aux précédentes approches sur une base de données publique, appelée Watch-n-Patch" [91], comme le montre la Table 3.1.

Références	Année	Méthodes	Caractéristiques	Taux de vrais positifs	Taux de faux négatifs
Nghiem et al. [88]	2012	SVM	HOG modifié	0.519	0.076
Chen et al. [89]	2016	LDA	Descripteur de tête circulaire	0.709	0.108
Ballotta et al. [90]	2017	CNN	Apprentissage profond	0.883	0.077
Ballotta et al. [91]	2018	CNN	Apprentissage profond	0.964	0.036

TABLE 3.1 – Performances des approches de détection de la tête appliquées sur la base de données Watch-n-Patch

3. 1.2.2 Analyse de la silhouette de la personne

La silhouette humaine, généralement décrite sous la forme d'une ellipse ou une boîte qui l'englobe, est extraite par des techniques de traitement d'images telles que la segmentation de l'avant plan ou la soustraction d'arrière-plan. Cette silhouette est alors communement modelisée sous la forme d'une eellipse ou une bite qui l'englobe. Les paraamètres de ce modèle, comme par exemple la hauteur, la largeur et l'angle de rotation de l'ellipse (Figure 3.5), fournissent un indice important sur le fait qu'une personne peut être debout ou couchée lors d'une chute. Vaidehi et al. [92] ont utilisé les caractéristiques statiques de l'ellipse englobante de la personne telles que les axes et l'angle d'inclinaison. Gasparrini et al. [93] ont proposé un système de détection des chutes basé sur la segmentation de la silhouette sur des images de profondeur. Ils ont identifié la chute par la distance entre la tête et le sol et ont suivi la personne à l'aide d'une caméra de profondeur pour détecter les chutes. Bian et al. [94], Zhao et al. [37] et Li et al. [95] ont proposé aussi une méthode de suivi des parties du corps humain à partir d'une caméra de profondeur. Le système développé par Bansal et al. [96] est fondé sur l'analyse et le suivi de la boîte englobante de la silhouette de la personne et sur la reconnaissance du squelette extrait à l'aide de capteur Kinect. Chua et *al.* ont utilisé une stratégie qui consiste à décomposer le corps humain en 3 parties, à ajuster ces parties sur la silhouette pbtenue après soustraction du fond, et à detecter la chute par analyse de ces 3 parties [97]. Cette représentation est

sensible à plusieurs mouvements, tel l'exemple d'une personne qui ramasse quelque chose par terre et se lève brusquement. L'étude faite par Abobakr et al. [98] a mis en évidence que la détection des chutes en suivant le squelette réduit les performances du traqueur vu que le suivi des articulations ne peut plus extraire le squelette lors d'un mouvement rapide comme la chute en raison de l'orientation du corps qui change d'une façon significative. Face à cette difficulté, Amini et al. [99] ont développé un modèle plus performant en ajoutant la position 3D de la tête qui a été calculée en projetant des lumières infrarouges sur les objets et en calculant le temps de réception de chaque faisceau par le récepteur infrarouge du capteur.

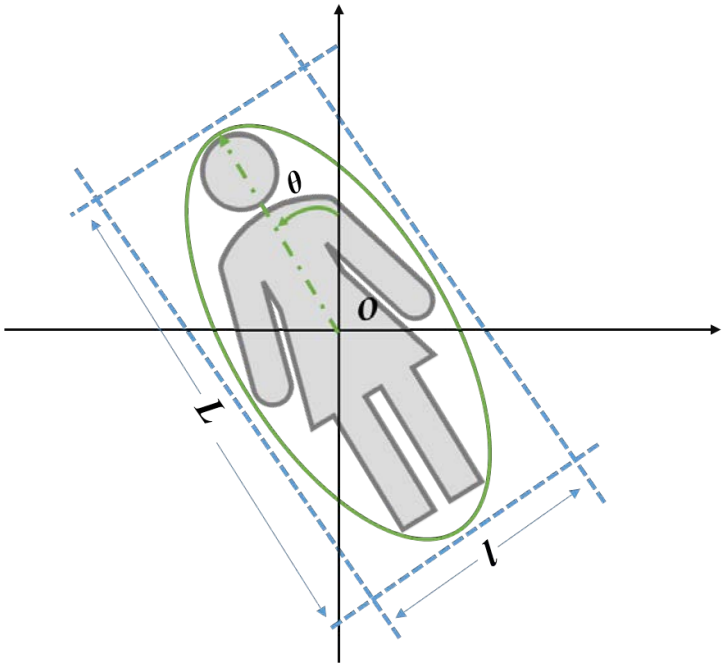


FIGURE 3.5 – Forme elliptique englobant la silhouette de la personne.

3. 1.2.3 Détection de l'inactivité de la personne

Certaines recherches ont étudié les mouvements de la silhouette de de la personne durant une séquence et ils ont considéré son inactivité comme indice de détection des chutes [100]. Selon d'autres [101], les chutes présentent des caractéristiques spécifiques qui peuvent être exploitées pour les détecter et les prévoir. Ces caractéristiques peuvent être une augmentation négative de l'accélération du corps due à un changement rapide

de position qui se termine par une période d'inactivité de la personne sur le sol ou un changement brusque au niveau de hauteur, de largeur et d'inclinaison de l'angle du corps. En général, il considère qu'une personne est tombée lorsqu'elle se trouve allongée dans une région définie à risque comme le sol et dans un état inactif. Cette technique est utilisée comme un indice pour la détection des chutes, comme présenté par Vaidehi et al. [92], Yu [102] et Rougier et al. [87]. Jansen et al. [84] ont également utilisé la silhouette et l'orientation de la personne pour analyser la chute en utilisant des images de profondeur. Le non-changement d'orientation du corps durant un certains laps de temps est utilisé pour identifier l'inactivité de la personne. Si l'inactivité est constatée et que la personne se trouve dans une zone définie comme à risque, une chute est détectée.

3. 2 Le mouvement d'une personne modélisé sous la forme d'un système dynamique

Dans le paragraphe précédent, le mouvement de la personne était considéré comme une suite temporelle de positions statiques d'un objet caractéristique. Or, il nous parait important de faire une analyse dynamique de ces mouvements dans la scène. Ceci passe par une modélisation des mouvements de la personne. Dans leur article, Fan et al. ont classé les modèles basés sur le mouvement de l'objet en deux catégories : Méthodes probabilistes et non probabilistes [35]. Le suivi probabiliste a été proposé pour tenir compte de l'incertitude dans une séquence. C'est un processus temporel qui estime l'état d'un système à partir de son état précédent et d'une série d'observations bruitées.Nous allons présenté 3 des modèles probabilistes les plus utilisés.

3. 2.1 Modèle de Markov caché

Un Hidden Markov Model -HMM- peut être adopté pour le suivi d'objets visuels grâce à sa capacité à gérer les dégradations introduites lors du processus d'acquisition. En effet, le modèle de Markov caché est composé de deux couches : une couche cachée, qui représente l'état du système et qui est décrit par un vecteur d'état x_t et une couche d'observations z_t, qui permet de mettre à jour l'état actuel à l'aide de l'état précédent (le temps est discrétisé). La Figure 3.6 montre une vue schématique du modèle. Le diagramme peut être mathématiquement exprimé comme suit :

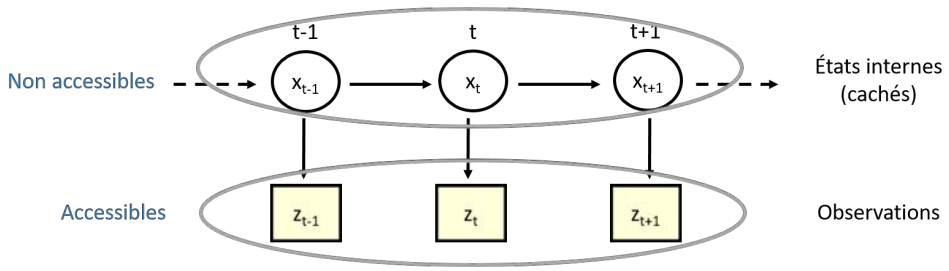


FIGURE 3.6 – Modèle de Markov caché.

$$x_t = f_t(x_{t-1}, v_{t-1}) \tag{3.5}$$

$$z_t = h_t(x_t, w_t) \tag{3.6}$$

où f_t et h_t sont des fonctions vectorielles ; on suppose qu'elles sont connues, éventuellement elles sont non linéaires et dépendantes du temps. Les fonctions dépendent des états x_{t-1} et x_t et des bruits de processus v_{t-1} et d'observation w_t. Le modèle de Markov caché s'appuie sur le filtrage bayésien récursif.
Le suivi des objets par cette approche bayésienne est modélisé comme un problème d'estimation de l'état actuel en se basant sur l'état précédent et les observations actuelles.

En se basant sur les équations (3.5) et (3.6), l'approche bayésienne sert à estimer correctement l'état x_t à partir de l'ensemble des observations $Z_t = z_1, z_2, ..., z_t$ disponibles. En d'autres termes, la récursion bayésienne examine la *densité a posteriori* $p(x_t|Z_t)$ pour estimer l'état de l'objet en utilisant la règle de Bayes. Supposons que nous observons Z_t qui est lié aux états cachés $X_t = x_1, x_2, ..., x_t$, le lien $p(Z_t|X_t)$ entre X_t et Z_t est connu ainsi que la probabilité de transition $p(X_t|X_{t-1})$.

Dubois et *al.* ont développé un système de reconnaissance des activités humaines, y compris la chute, à l'aide d'une caméra RGB-D en utilisant un modèle de Markov caché [103]. D'autres systèmes de détection de chute pour des applications in-door [104] ont été modélisés à l'aide d'un modèle HMM basé sur la vitesse verticale, la variation de surface et la hauteur d'une personne. Dans cette étude, les caméras étaient positionnées de manière à fournir une vue de dessus.

3. 2.2 Filtre de Kalman

Le filtre de Kalman, basé sur une distribution gaussienne, est un outil récursif. Appliqué au problème du suivi de personnes, le filtre de Kalman sert à estimer la position d'une personne à l'instant $t+1$ à partir de celle estimée à l'instant t. Pour chaque nouvelle position à estimer, le filtre de Kalman se déroule en 2 étapes de façon cyclique (fig. 3.7).

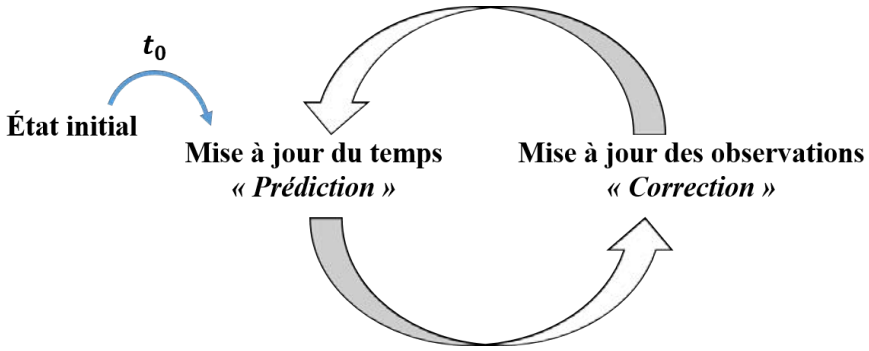


FIGURE 3.7 – Processus de filtre de Kalman.

Soit un système dynamique composé de deux vecteurs aléatoires X et Z, où X est le processus d'état, représentant ce que l'on cherche et Z le vecteur d'observation. Dans le cadre du suivi d'un objet, X peut représenter le centre du rectangle englobant de cet objet, et Z les différentes mesures (hauteur, largeur, surface de ce rectangle, etc). Le but de méthodes dites directes est de trouver la transformation $Z = f(X)$ qui permet de faire correspondre une observation Z avec une prédiction de l'état X. Les méthodes probabilistes cherchent à estimer la loi de probabilité a posteriori de cet état. Ce processus cyclique consiste à prédire la position de la prochaine itération (phase de *prédiction*), puis à ajuster la position actuelle en fonction de la position prédite et de celle qui a été mesurée (phase de *correction*) [105].
Le modèle de Kalman est basé sur l'hypothèse que l'état réel X au temps t dépend de l'état précédent t 1 suivant l'équation :

$$X_t = F_t X_{t-1} + w_t \tag{3.7}$$

où F_t est la matrice de transition appliquée à l'état précédent et w_t est le bruit d'état. A l'instant t, une observation est réalisée, selon l'équation :

$$Z_t = H_t X_t + u_t \tag{3.8}$$

où H_t est la matrice d'observation et u_t le bruit gaussien.

Des extensions de ce filtre ont été créées pour les modèles non linéaires [106, 107, 108], qui sont utilisées pour le suivi de personnes [109, 110, 111] et de véhicules [112, 113]. Le filtre de Kalman a montré de bonnes performances. Néanmoins, ce filtre n'est pas aussi efficace lorsqu'il s'agit de trajectoires aléatoires de personnes et surtout d'objets déformables dans le cas d'occultations ou de changement brusque de direction [114].

3. 2.3 Filtre particulaire

Contrairement à l'algorithme de Kalman, les filtres particulaires, basés sur l'algorithme CONDENSATION (Conditional Density Propagation) [115], sont utilisés pour des systèmes non linéaires avec des modèles non Gaussiens d'observation. Ainsi, il est bien adapté pour suivre une trajectoire avec des brusques changement de direction. Il est basé sur le principe de simulation de Monte Carlo et il a pour but d'estimer récursivement la densité de probabilité *a posteriori* $p(X_t|Z_{1:t})$ du vecteur d'état X_t à l'instant t conditionné sur l'ensemble de mesures $Z_{1:t} = Z_1, ..., Z_t$.Á chaque instant t, on approxime la densité $p(X_t|Z_{1:t})$, grâce à un ensemble de "N particules" s_t^i. Chaque particule représente une hypothèse de l'état avec son poids correspondant à la probabilité conditionnelle $\pi^n = p(z_t|s_t^n)$. Ces particules seront corrigées en fonction de leur cohérence avec les observations et leur poids. Grâce à cette correction, les particules évoluent dans le temps et sont échantillonnées selon une fonction d'importance afin d'explorer les zones pertinentes de l'espace d'état. Une particule d'indice i est représentée par le couple (X_t^i, π_t^i) [114].

L'algorithme CONDENSATION est une version de filtrage particulaire, utilisé dans le domaine de suivi visuel. Cet algorithme est composé de trois étapes principales (Figure 3.8) :

- **Échantillonnage** : Un nouvel ensemble de particules est construit aléatoirement en favorisant les particules de poids les plus importants de l'ensemble précédent.
- **Prédiction** : Chaque particule est propagée à l'aide d'un modèle dynamique stochastique.
- **Correction** : Un poids est calculé pour chaque particule du nouvel ensemble.

Une fois que tous les particules sont à jour, le résultat du suivi est estimé par la moyenne pondérée ou la particule de poids maximal de cet ensemble de particules. Les filtres particulaires ont été utilisés pour suivre des personnes [116], mais aussi des parties du corps telles que la tête ou les mains [115, 117, 118]. Cette méthode de suivi semble très prometteuse et est utilisée dans de nombreux travaux récents. Elle donne d'excellents

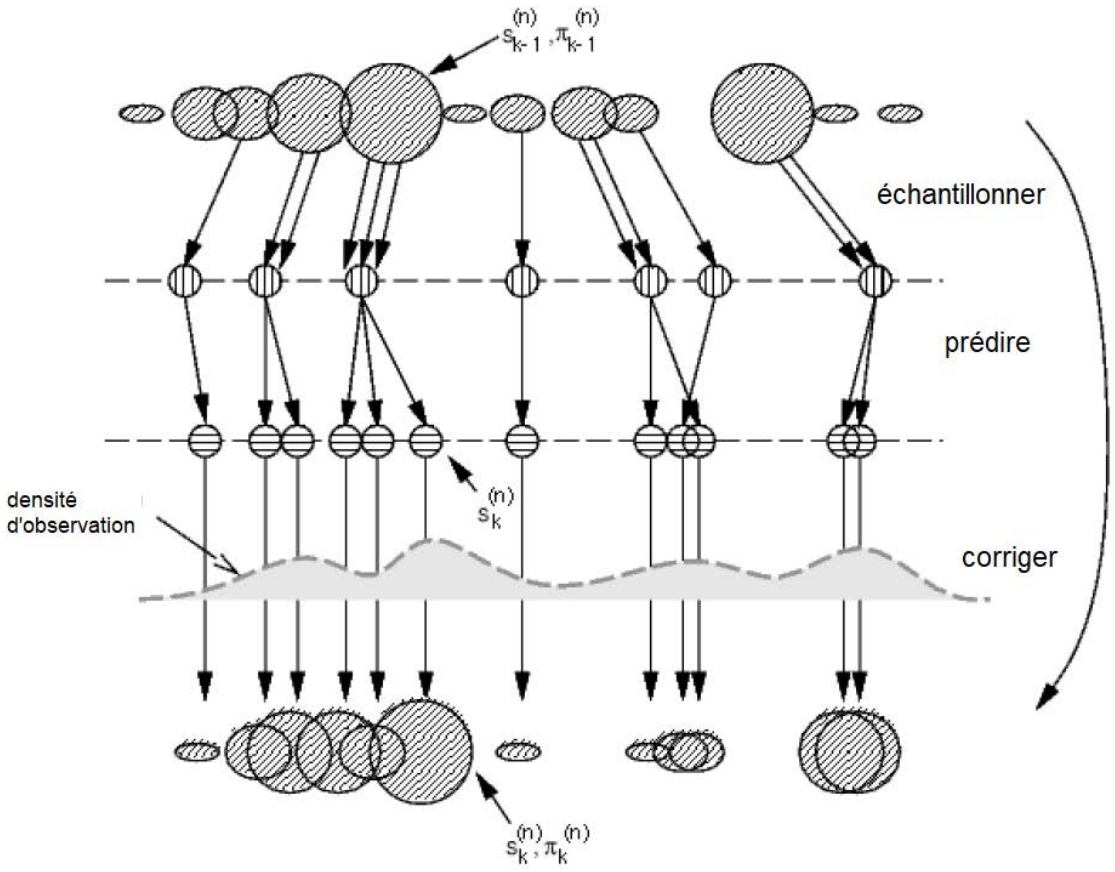


FIGURE 3.8 – Une itération de filtrage particulaire [115].

résultats à condition d'avoir suffisamment de particules pour avoir un bon maximum de vraisemblance pour l'état estimé [114].

3. 3 Conclusion

L'un de nos objectifs, comme nous l'avons mentionné, est de développer un système peu coûteux. En d'autres termes, le système doit pouvoir être installé sur une machine à bas coût. Les méthodes comme l'extraction du squelette, par exemple, sont des méthodes très précises mais coûteuses en temps de calcul. Une des méthodes les plus classiques pour segmenter la personne, et que nous avons sélectionnée pour développer notre système de suivi, est la soustraction du fond. Cette méthode est considérée comme l'une des moins robustes pour les images couleur. Mais, nous traitons les images de profondeur et thermique. En effet, la représentation du fond peut être construite sur l'information de profondeur de chaque pixel, ce qui rend cette méthode ainsi plus robuste, car n'étant plus dépendante des changements de lumière. De plus, nous avons basé notre modèle de suivi, qui est détaillé dans le chapitre suivant, sur le centre de la tête de la personne. Afin de décrire le mouvement nous avons également choisi d'utiliser le filtre particulaire car il semble bien adapté à suivre une trajectoire avec des brusques changement de direction comme nous pouvons trouver dans une chute.

Chapitre 4

Suivi de la tête par filtrage particulaire dans le contexte de détection de la chute sur des images de profondeur

---

Dans le chapitre précédent, nous avons détaillé les différentes méthodes de suivi de personnes. Dans ce chapitre, nous présentons, tout d'abord, notre algorithme de suivi basé sur le filtrage particulaire. Ensuite, nous testons cet algorithme sur des images de profondeur. Enfin nous comparons les résultats de cet algorithme avec ceux de la segmentation seule.

Introduction

Dans cette partie, nous présentons les différentes étapes de notre modèle de suivi de la personne. Ces étapes sont illustrées dans la Figure 4.1. Elles sont au nombre de deux : (i) Segmentation de la silhouette de la personne et détection de sa tête (ii) Suivi de la représentation de la personne. Ces deux étapes vont structurer notre chapitre, la première partie du chapitre décrit les différentes étapes d'extraction de la silhouette dans la scène et d'estimation de la position de la tête de la personne. La deuxième partie contient l'algorithme de suivi de la région d'intérêt déterminée dans l'étape de segmentation.
Le principe de l'algorithme est de suivre une personne à travers des paramètres efficaces et simples à extraire. Nous avons choisi de suivre, non la personne (sa silhouette) directement, mais plutôt la tête puisque cette dernière est :

— La partie la plus haute du corps
— Peu déformable
— Moins sujette à occultation

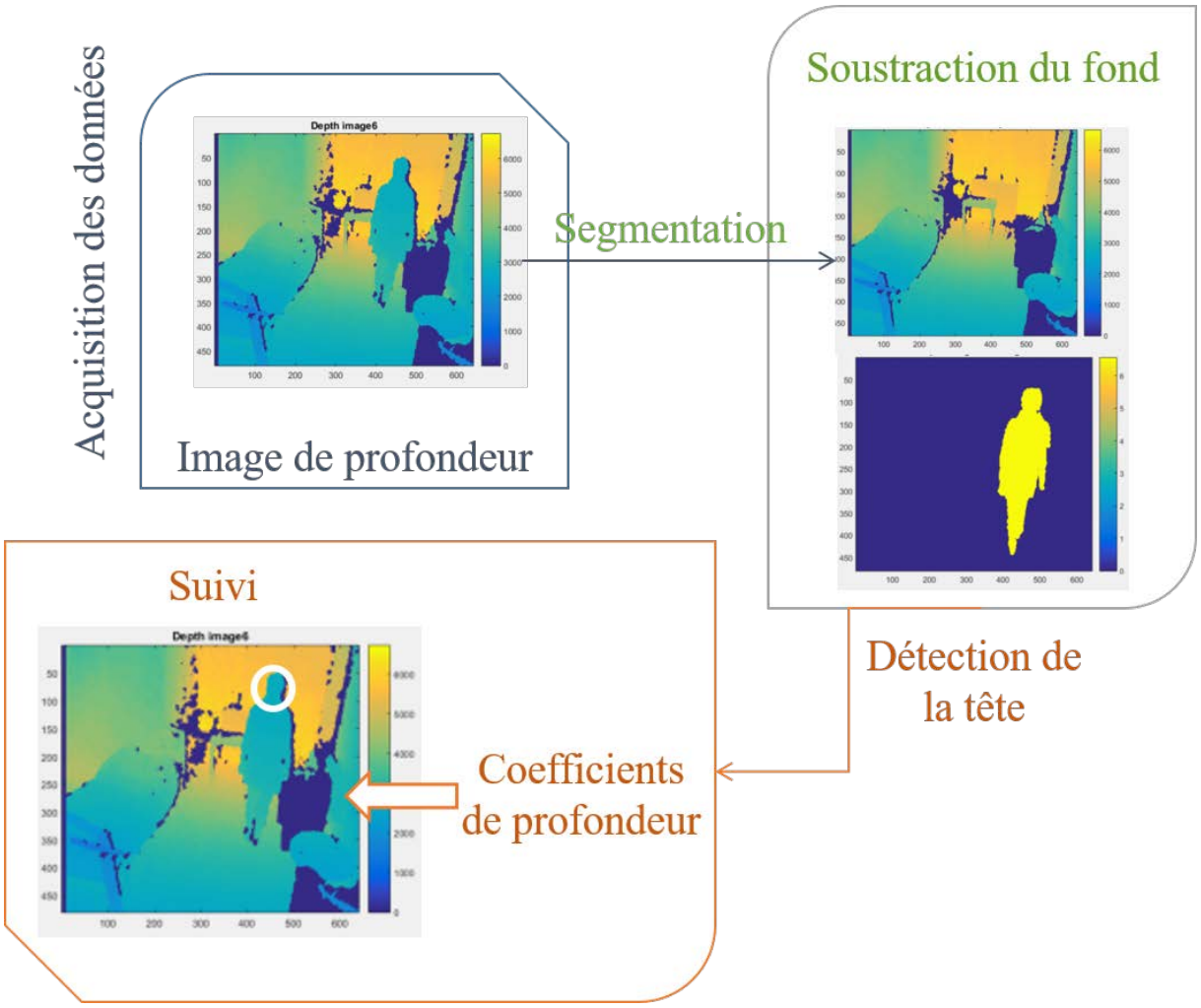


FIGURE 4.1 – Illustration de notre algorithme de suivi de la tête d'une personne.

— Une forme simple à approximer par une ellipse
— Sa vitesse verticale est significative durant une chute

4. 1 Segmentation de la silhouette de la personne sur des images de profondeur

Dans cette section, nous présentons la segmentation de la silhouette de la personne à partir des images de profondeur fournies par la Kinect. Pour ce faire, nous suivons l'organigramme illustré dans la Figure 4.2. Tout d'abord, nous commençons par créer la carte de référence, c'est-à-dire sans silhouette. Puis nous détectons le premier plan en soustrayant cette carte de référence de chaque image. Ensuite nous segmentons la silhouette de la personne en la délimitant par une ellipse. Enfin nous extrayons le centre de la tête en se basant sur cette ellipse et nous approximons la tête par une ellipse de même centre que la tête.

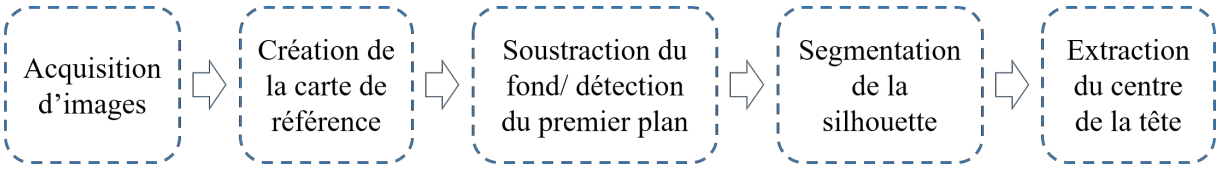


FIGURE 4.2 – Processus d'extraction du centre de la tête.

4. 1.1 Création de la carte de référence

La création de la carte de référence représente la première étape de segmentation. Cette étape sert à identifier l'arrière-plan ou le fond de la scène représenté par les pixels fixes de l'image. La technique utilisée pour apprendre ces pixels est basée sur la moyenne et l'écart type des N premières images vides qui ne contiennent aucun objet en mouvement (voir Figure 4.3), cette carte est appellée BG (BackGround).

4. 1.2 Segmentation

Dans cette étape, nous cherchons à identifier les pixels mobiles de l'image qui donc définissent la personne. Pour chaque nouvelle image I_t, le fond est soustrait dans un premier temps. Sur l'image de différences un pixel $p(x, y)$ appartient au premier plan tant

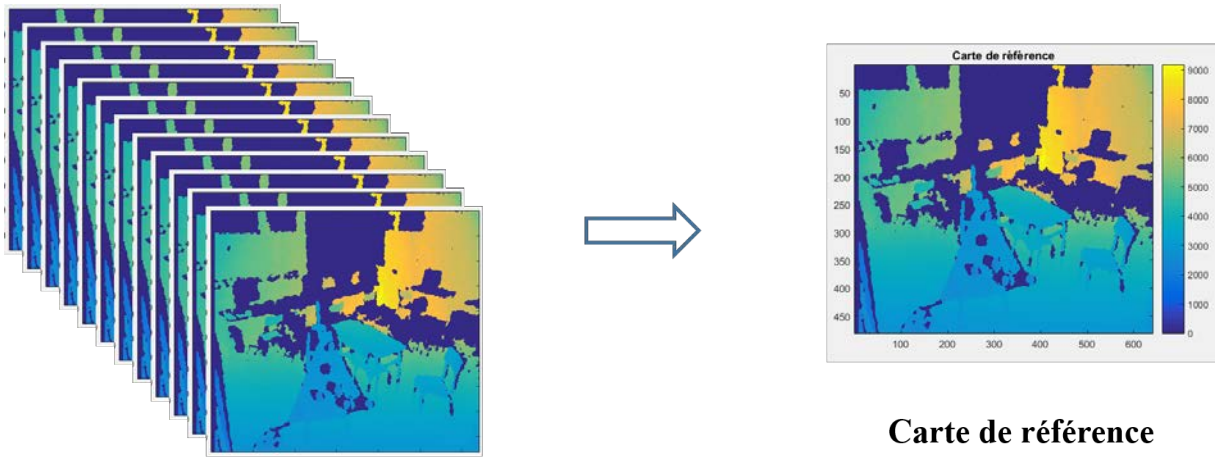


FIGURE 4.3 – Création de la carte de référence.

qu'il est au-dessus d'un seuil fixé en fonction de la variance de la carte de référence. Dans le cas contraire, le pixel est classé comme arrière-plan :

$$\begin{aligned} Si\ |I_t(x,y) - BG(x,y)| > 2\sigma(x,y) \\ alors\ I_t(x,y) \in Premier\ plan \\ sinon\ I_t(x,y) \in Arrire\ plan \end{aligned} \tag{4.1}$$

avec $\sigma(x,y)$ l'écart type de $BG(x,y)$.

La soustraction permet bien d'estimer les objets en mouvement mais également des pixels "instables" qui peuvent aléatoirement se détacher du fond du fait de la technique de mesure de la profondeur. Pour éliminer ce bruit ou renforcer les clusters de pixels, deux opérations morphologiques sont utilisées : érosion et dilatation. L'opération d'érosion ellimine les pixels isolés. Elle sert à reduire le bruit. L'opération de dilatation remplit la majorité des trous pour retrouver la silhouette originale [119].

L'étape de segmentation est ainsi capable de détecter toutes les régions qui contiennent des objets en mouvement dans la scène. Nous faisons l'hypothèse qu'il n'y a qu'une personne. Donc nous choisissons la région ayant le nombre de points mobiles le plus important. Finalement, nous encadrons cette région par une ellipse.

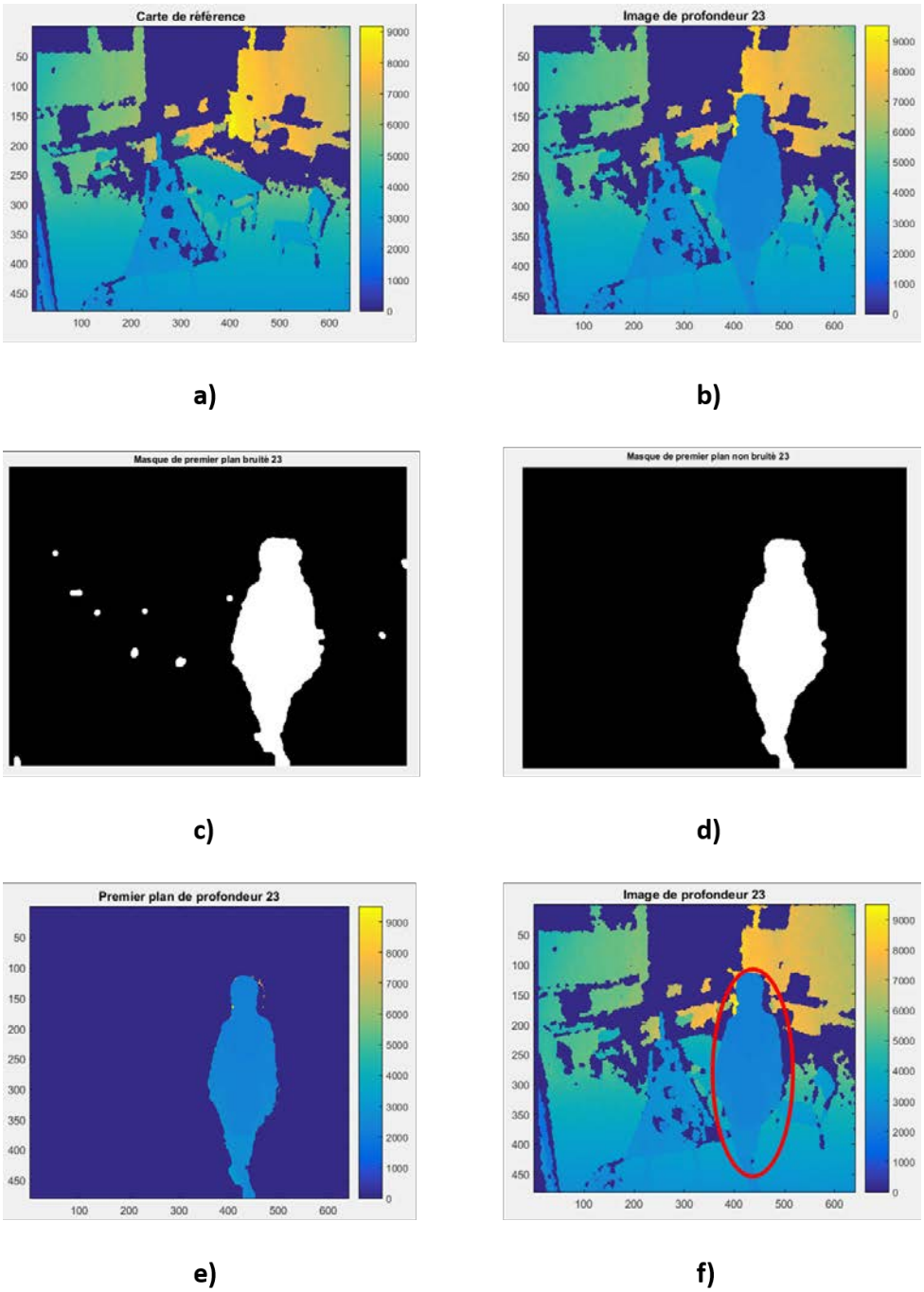


FIGURE 4.4 – Segmentation de la silhouette : a) carte de référence, b) image actuelle, c) détection du premier plan, d) suppression du bruit e) extraction de la silhouette et f) détection de la silhouette sur l'image actuelle.

La Figure 4.4 illustre chacune des étapes définies ci-dessous à l'aide d'un exemple. Nous constatons l'importance de la phase de segmentation pour suivre correctement la personne sur des images de profondeur qui ne sont pas de haute qualité.

4. 1.3 Extraction du centre de la tête

La tête ne peut pas être segmentée directement à partir des informations contenues dans l'image. Nous nous servons de l'ellipse ajustée à la silhouette pour estimer le centre de la tête sur les images de profondeur. La tête représente environ un sixième de la longueur totale du corps. Alors, nous avons fixé le centre de la tête H à un sixième du demi-grand axe CT par rapport à la partie supérieure du grand axe (voir Figure 4.5). De plus nous avons modélisé la tête par une ellipse ayant la même orientation que celle de la silhouette mais réduite à un sixième de sa taille. Cette modélisation nous permet d'évaluer l'impact de différents paramètres représentant la tête.

$$\|\overrightarrow{HT}\| = \frac{1}{6}\|\overrightarrow{CT}\| \tag{4.2}$$

où C est le centre de l'ellipse de la silhouette, T est le point supérieur du grand axe et BT est le grand axe de la tête sachant que le rapport entre le grand axe et le petit axe est fixé à 1.2 [120].
Par contre, la seule utilisation du processus de segmentation n'est pas une méthode suffisamment robuste pour suivre la position de la tête. D'une part la tête n'est pas toujours l'objet le plus haut de la silhouette et d'autre part, une fausse silhouette, et donc une fausse tête peut être détectée comme le montre la Figure 4.6. L'utilisation d'une méthode de suivi est donc nécessaire.

4. 2 Suivi de la personne à partir des images de profondeur

L'objectif de cette étape est d'estimer la position de la tête à chaque nouvelle image en prenant en considération le dernier état de la tête. Nous avons donc choisi une méthode séquentielle de Monte Carlo qui est la méthode du filtre particulaire (PF).

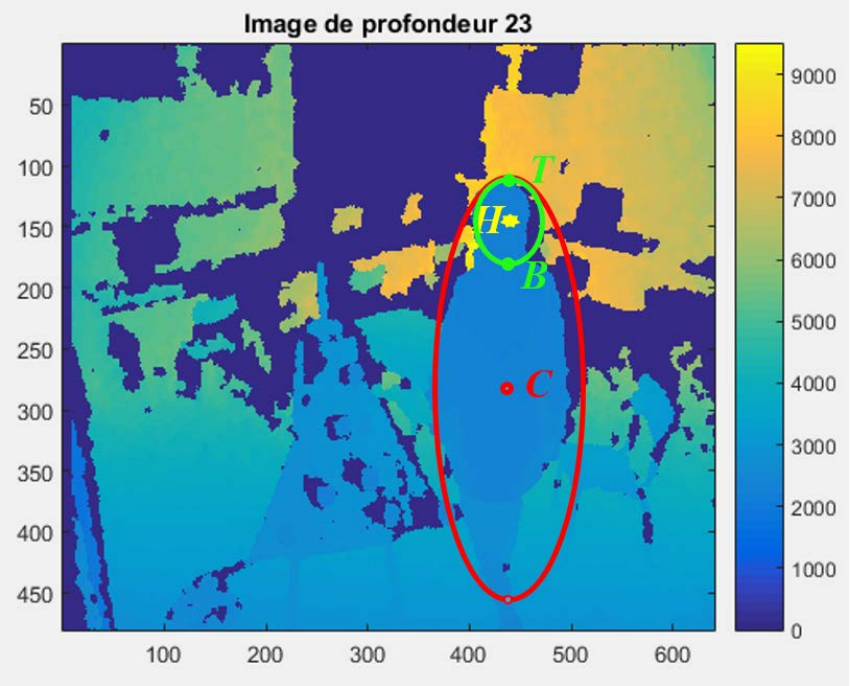


FIGURE 4.5 – Position de la tête en fonction de la silhouette de la personne. Les ellipses rouge et verte englobent respectivement la silhouette et la tête

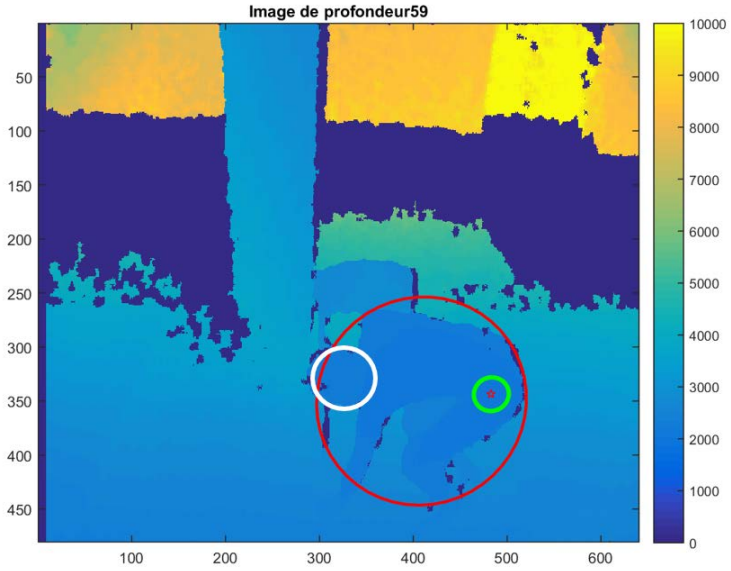


FIGURE 4.6 – Estimation de la position de la tête (ellipse verte) en fonction de la segmentation de la silhouette de la personne (ellipse rouge) et comparaison avec la vérité terrain (ellipse blanche).

4. 2.1 Filtre particulaire

Dans le chapitre précédent, nous avons montré l'importance du filtrage particulaire qui traite notamment des systèmes non linéaires. L'objectif d'un filtre particulaire est d'estimer la densité *a posteriori* de variables d'état compte tenu de variables d'observation. Dans notre cas, pour une image à l'instant t, un vecteur d'état x_t est composé par la position de la tête H, et d'autres paramètres permettant de mieux caractériser la tête comme la taille L et/ou son orientation θ (en fait, la taille et/ou l'orientation de l'ellipse englobante). L'équation 4.3 présente un exemple de vecteur d'état :

$$x_t(S1) = (x_H, y_H, L, \theta)^T \tag{4.3}$$

La méthode *PF* cherche à estimer le vecteur d'état caché x_t à partir du vecteur d'état précédent x_{t-1} et des vecteurs d'observations d'information de profondeur $Z_t = \{z_1, \cdots, z_t\}$.

PF utilise un échantillon de N particules $S_t = \left\{S_t^1, \cdots, S_t^N\right\}$ pour obtenir une approximation de la probabilité conditionnelle $p(x_t/Z_t)$. Chaque particule S_t^n peut être considérée comme une hypothèse sur x_t. Elle est pondérée par un poids d'importance π_n qui est lié à la probabilité *a posteriori*. L'ensemble des poids d'importance est normalisé $\sum_{n=1}^{N} \pi_n = 1$.

La première étape du suivi est l'initialisation des particules. Pour cela nous partons d'un vecteur d'état initial correspondant à la tête segmentée sur la première image de la séquence. Les N particules sont alors créées en ajoutant une modification aléatoire des paramètres du vecteur d'état initial suivant une loi Gaussienne. Le poids de chaque particule est initialisé à $\frac{1}{N}$.

Après quelques itérations, certaines particules peuvent avoir un poids proche de zéro (cf. section 3.2.3). Alors, l'estimation de la nouvelle position n'est plus basée sur l'état de N particules mais sur un nombre plus faible de particules qui peut entraîner la divergence du filtre. Afin de prévenir ce problème de "dégénérescence des particules", les particules sont ré-échantillonnées en un nouvel ensemble de N particules S_{t+1} à chaque nouvelle image $t+1$. De plus, nous avons choisi de mettre à jour les poids des particules à partir de la probabilité des nouvelles observations (détaillé dans la section 4.2.2). Notre première contribution permet d'éviter la dégénérescence rapide des particules en remplaçant la dépendance des poids des particules d'une image à une autre par une distribution gaussienne qui dépend uniquement des observations à l'instant actuel.

Le ré-échantillonnage a pour objectif de dupliquer les particules dont le poids est suffisamment fort et d'éliminer les particules dont le poids est faible afin de ne conserver que les particules les plus significatives. Pour ce faire, nous avons appliqué l'algorithme suivant :

Algorithm 1 : ré-échantillonner (S_t)

for $i = 1 : N$ **do**
  $u \sim \mathcal{U}[0, 1]$;
  $j \leftarrow 1$;
  while $w_1 + ... + w_j < u$ **do**
    $j \leftarrow j + 1$;
  end
  $newS_t^i = S_t^j$
end

Dans un second temps, les particules sont propagées en fonction de l'équation de prédiction suivante :

$$S_{t+1}^n = A * S_t^n + w_t \tag{4.4}$$

où A représente la matrice du modèle de transition et w_t est un bruit gaussien.

Dans un troisième temps, les poids sont alors remis à jour en fonction des nouvelles observations (coefficients) que nous appelons new_{obs}. new_{obs} est une combinaison linéaire de coefficients issus de l'image courante à $t+1$. La constitution de new_{obs} est un des points critiques de la méthode, elle sera décrite dans la section 4.2.4.

Au final, l'avantage de *PF* est qu'il ne considère pas qu'un seul vecteur d'état mais N vecteurs d'état (un par particule). Par contre, à chaque image, la position de la tête doit être estimée à partir de ces N vecteurs d'état. Nous avons essayé deux types de modèles pour cette estimation : un premier modèle, appelé *AM*, qui considère que x_t est la moyenne pondérée des N particules et le second modèle, appelé *MM*, qui ne garde que la particule ayant le poids maximum.

La suite des étapes, comportant la segmentation et le suivi par *PF*, est illustrée dans la Figure 4.7.

Ainsi, les étapes de l'algorithme itératif de suivi *PF* sont :

1. **Initialisation** : générer un échantillon de N particules $S_1 = (S_1^1, ..., S_1^N)$ en se basant sur le vecteur d'état initiale x_1, et initialiser le poids de chaque particule

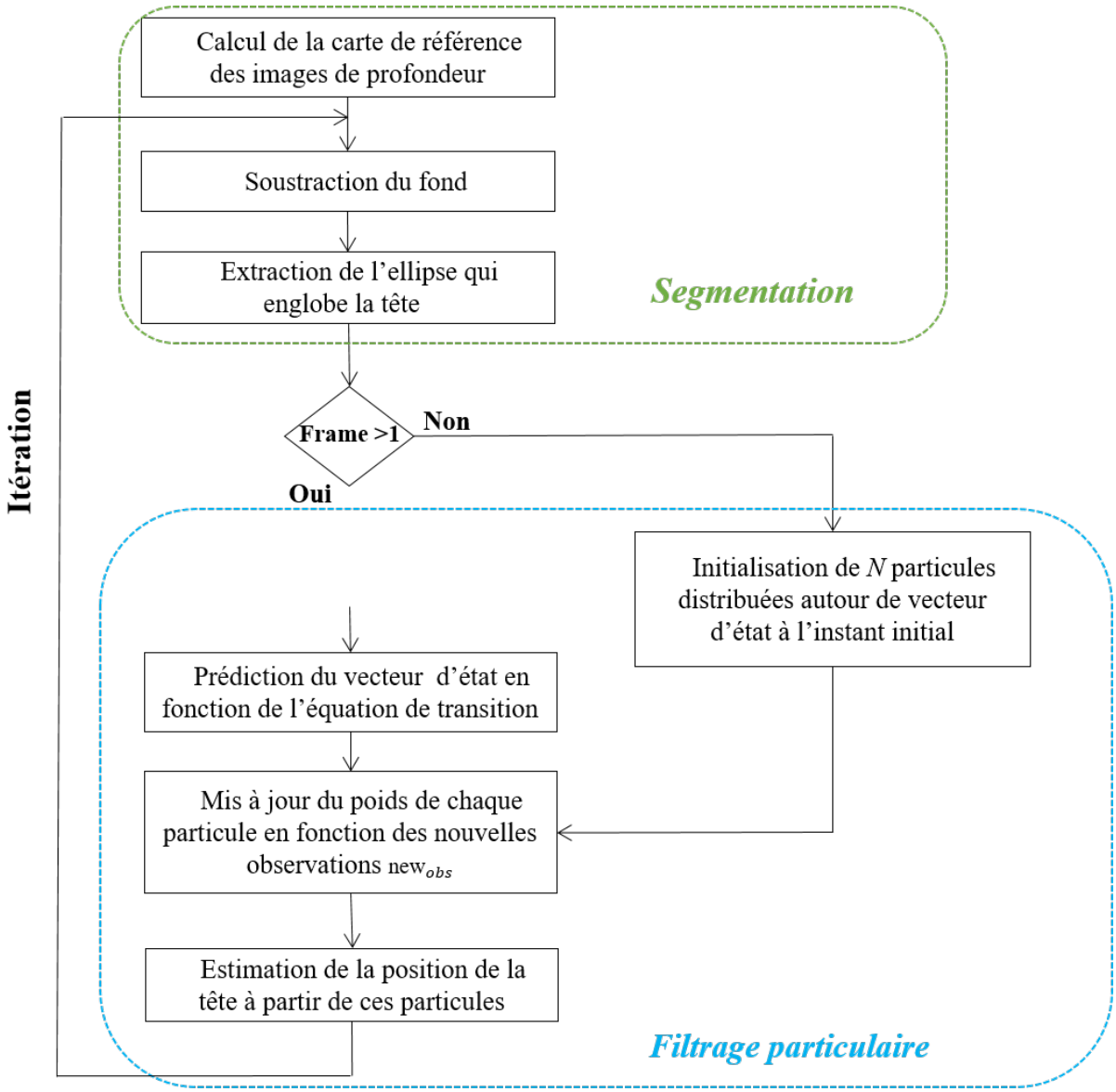


FIGURE 4.7 – Algorithme itératif de suivi basé sur le filtrage particulaire.

par :

$$\pi_1(n) = \frac{1}{N} \tag{4.5}$$

2. **Ré-échantillonnage** : ré-échantillonner l'ensemble des particules selon l'algorithme 1.

3. **Prédiction** : propager les particules selon le modèle de prédiction (éq. 4.4) pour prédire le vecteur d'état x_t.

4. **Mise à jour** : mettre à jour l'état prédit en calculant les poids des particules $\pi_1(n)$ en fonction des vecteurs d'observation (cf. paragraphe 4.2.2). Ensuite, normaliser le poids comme suit :
$$\pi_t(n) = \frac{\pi_t(n)}{\sum_{k=1}^{N} \pi_t(k)} \tag{4.6}$$
et retour à l'étape 2.

4. 2.2 Mise à jour des poids

La mise à jour de l'état prédit est l'étape principale du filtrage particulaire. Cette étape permet de corriger les poids des particules et elle est spécifique à chaque application (voir [117] pour un exemple de traitement d'images couleur).
Le poids d'une particule est défini par l'équation 4.7 :

$$\pi_t(n) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{new_{obs}(n)}{2\sigma^2}} \tag{4.7}$$

où σ est théoriquement l'écart-type de la combinaison des coefficients. Cependant, nous avons choisi une valeur constante de σ pour des raisons de coût de calcul. Nous avons testé plusieurs valeurs et nous avons choisi celle qui donnait le meilleur résultat. new_{obs} apporte l'information de l'image courante. C'est une combinaison linéaire de coefficients extraits de l'image de profondeur. Les coefficients choisis et la façon de les estimer sont décrits dans la section suivante.

4. 2.3 Choix des coefficients de profondeur

Nous avons choisi trois coefficients pour l'image de profondeur : C_F^P (Coefficient d'avant-plan : Foreground coefficient), C_G^P (Coefficient de gradient de profondeur) et C_D^P (Coefficient de distance de profondeur).

4. 2.3.1 Coefficient d'avant-plan

Le coefficient d'avant-plan C_F^P mesure en quelque sorte la concordance de la tête par rapport à l'image de l'ellipse portée par chaque particule $s_T^i(L, l)$ où L est le grand axe et l le petit axe [121].
Dans la Figure 6.11, l'ellipse centrale (en bleu), est calculée à partir du vecteur d'état de la particule considérée, le coefficient d'avant-plan C_F^P est calculé à partir de cette ellipse et deux autres ellipses calculées comme suit : l'ellipse externe (en vert) $s_{ext}(5L/4, 5l/4)$ et l'ellipse interne (en jaune) $s_{int}(3L/4, 3l/4)$.

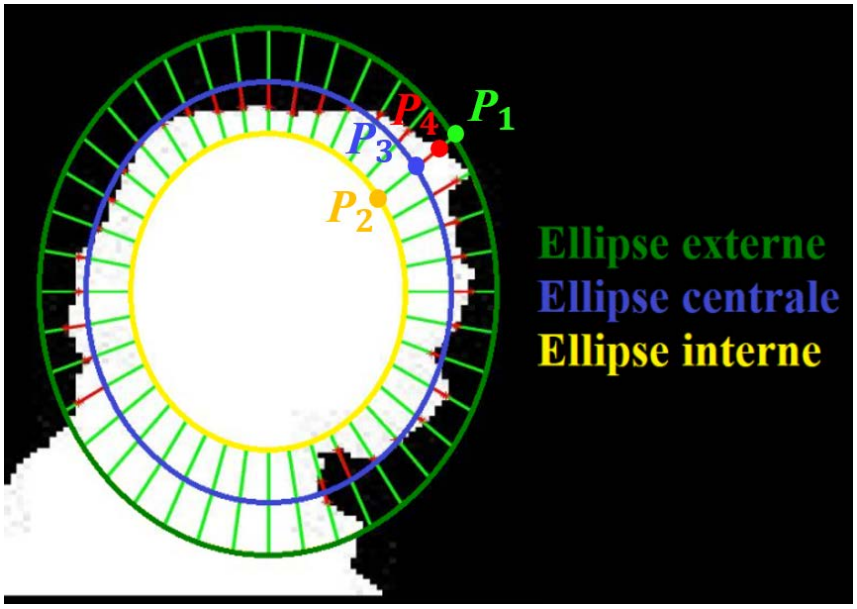


FIGURE 4.8 – Représentation du coefficient d'avant-plan.

Ces ellipses sont reliées par des segments. On note P_1^n le point d'intersection de l'ellipse externe s_{ext} avec le n ème segment et P_2^n le point d'intersection de l'ellipse interne s_{int} avec le n ème segment. Pour ce segment n nous estimons la longueur $D(n)$ du demi segment par :

$$D(n) = \frac{1}{2}\|\overrightarrow{P_1^n P_2^n}\| \tag{4.8}$$

Si l'on considère maintenant le point P_3^n, point d'intersection du segment avec l'ellipse centrale, et P_4^n, l'intersection du segment avec le contour calculé après l'étape de segmentation, nous pouvons calculer la distance du point du contour par rapport à l'ellipse centrale par :

$$d(n) = \left\|\overrightarrow{P_3^n P_4^n}\right\| \tag{4.9}$$

Le coefficient C_F^P est alors défini par :

$$C_F = \frac{1}{N}\sum_{i=1}^{n}\frac{(D(i) - d(i))}{(D(i))} \in [0, 1] \tag{4.10}$$

Si l'ellipse centrale épouse bien la forme de la tête issue de la segmentation, C_F^P s'approche de 1. Dans le cas opposé, C_F^P tend vers 0 (voir l'exemple dans la Figure 4.9).

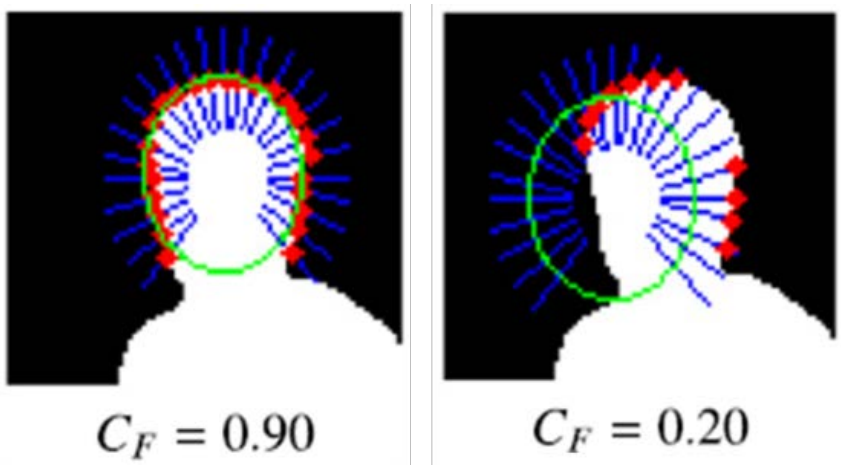


FIGURE 4.9 – Exemple de calcul du coefficient d'avant-plan [121].

4. 2.3.2 Coefficient de gradient

Le coefficient de gradient C_G^P mesure la somme des gradients autour du contour de l'ellipse de la particule, c'est-à-dire qu'il faut calculer la moyenne du gradient orthogonal à l'ellipse centrale $s_T^i(L, l)$ [122] qui est égale à :

$$C_G = \frac{1}{N}\sum_{i=1}^{N}(n(i).g_s(i))^- \tag{4.11}$$

$$(.)^- = |min(.,0)| \tag{4.12}$$

4. 2.3.3 Coefficient de distance

Le coefficient de distance C_D^P est la distance entre le centre de l'ellipse de la particule et le centre de la tête issu de la segmentation.

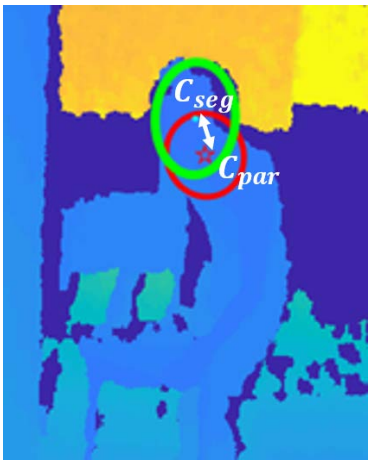


FIGURE 4.10 – Coefficient de distance. L'ellipse de la tête estimée par le suivi est en rouge et celle de la vérité terrain est en vert. La distance entre le centre de la tête estimé et la vérité terrain est le coefficient de distance.

4. 2.4 Résultats du suivi

Dans cette section, nous présentons dans un premier temps l'implémentation de la méthode de suivi ainsi que la base de données de tests et les critères d'évaluation. La deuxième partie contient les résultats que nous avons obtenus, suivis d'une discussion.

4. 2.4.1 Implémentation

La méthode décrite ci-dessus a été implémentée en Matlab et testée sur un ordinateur portable qui fonctionne sur Windows 10 64 bits, avec un processeur Intel Core I7 à 2.6

GHz et 16Go de RAM. Pour tous les tests, nous avons fixé la fréquence d'acquisition à 8 Hz.

Dans cette partie, les méthodes de suivi décrites dans les sections 4.1 et 4.2 sont testées sur des images de profondeur uniquement.Nous allons ainsi comparer les performances des différentes variantes de notre méthode de suivi par rapport à une simple segmentation de la tête selon la méthode décrite dans la section 4.1.2. Nous allons donc commencer par comparer les différentes compositions de vecteur d'état et leurs influences sur le modèle de suivi. Ensuite, nous allons étudier l'impact de chaque coefficient sur la mise à jour des poids ainsi que le suivi de la tête. Finalement, nous allons évaluer les performances des modèles AM et MM.

4. 2.4.2 Base de tests

Dans cette section, nous avons évalué les performances de l'algorithme proposé sur une base de tests constituée dans trois environnements différents : une salle dans le Living Lab ActivAgeing, une salle de Tp et un appartement à l'ECAM (voir section 2.4.2).
Cette base est constituée par différentes séquences contenant une seule personne à chaque fois. Elle contient plusieurs activités normales (marcher, s'allonger, manger, s'assoir, etc) et anormales (occultation, chute). Afin de diversifier les données nous avons fait jouer ces séquences par sept personnes de sexe, d'âge et de taille différents. L'expérience consistait à effectuer des activités de la vie quotidienne, notamment : se tenir debout, s'asseoir, marcher et se coucher. La caméra Kinect était fixée au plafond à une hauteur de 2.7 m. Pour évaluer ce modèle, nous l'avons comparé avec les résultats de la segmentation seule. La vérité terrain (GT : Ground truth en anglais) a été établie manuellement en identifiant la tête par une ellipse à chaque image.

4. 2.4.3 Critères d'évaluation

Afin d'évaluer la justesse du suivi, nous avons utilisé deux mesures quantitatives comme critères d'évaluation :

- **La courbe de précision** : dans notre cas, la précision est l'erreur de localisation du centre, définie par la distance euclidienne moyenne entre les positions de la tête détectées par l'algorithme de suivi et ceux de la vérité de terrain. Une première mesure de la performance globale de l'algorithme de suivi pourrait être l'erreur moyenne de localisation du centre de la tête sur toutes les images d'une séquence. Cependant, lorsque l'algorithme de suivi perd la cible, la valeur de l'erreur moyenne peut ne pas mesurer correctement les

performances du suivi [123]. C'est pour cela que nous avons adopté la courbe de précision pour mesurer la performance globale de suivi. Cette courbe a comme abscisses une distance et en ordonnées le pourcentage d'images de la séquence dont la position estimée se situe en dessous de cette distance par rapport à la vérité terrain. Cette méthode d'évaluation est déterminée par rapport à l'aire sous la courbe (AUC Area Under Curve en anglais). Plus cette aire est grande plus l'algorithme de suivi est performant [123], comme le montre la Figure 4.11.a.

- **La courbe de succès** : c'est le score de chevauchement des boîtes englobantes détectées par l'algorithme de suivi et de la vérité de terrain, respectivement [123]. Dans notre cas, nous avons choisi de calculer le score de chevauchement des ellipses, puisque ces dernières sont plus précises. Le score est défini par :

$$S = \frac{|T \cup G|}{|T \cap G|} \tag{4.13}$$

où $T \cup G$ et $T \cap G$ représentent respectivement l'union et l'intersection de deux ellipses et $|.|$ désigne le nombre de pixels dans la région. Ce score varie de 0 (pas de chevauchement) à 1 (ellipses parfaitement superposées).

Pour mesurer la performance d'une séquence d'images, nous traçons une courbe de Succès.

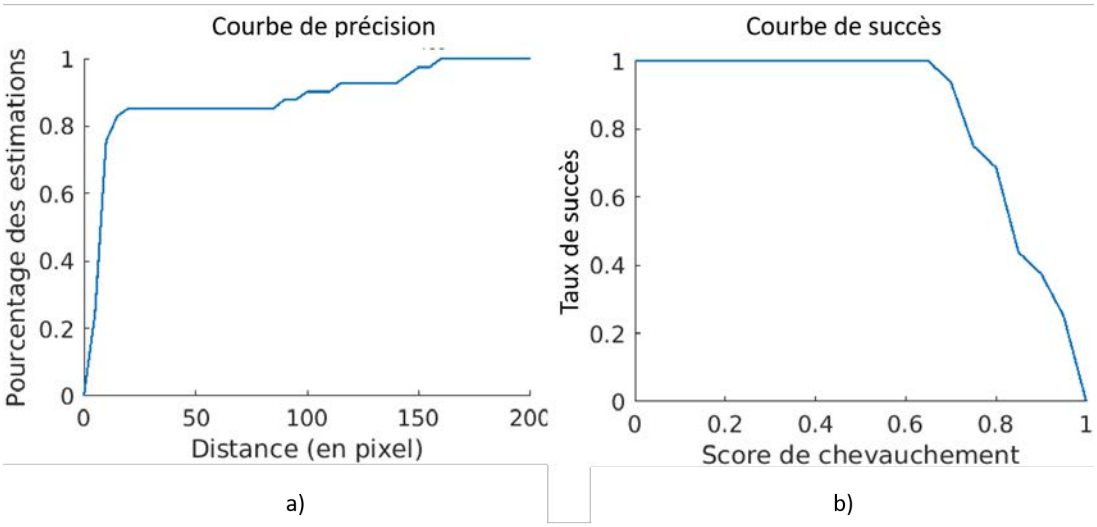


FIGURE 4.11 – Courbes d'évaluation de modèle de suivi : a) courbe de précision et b) courbe de succès

Cette courbe a pour abscisses un seuil de taux de chevauchement (entre 0 et 1) et en ordonnées le nombre d'images dont le score de chevauchement S est supérieur à ce seuil.

La performance peut alors être mesurée par la valeur de taux de succès à un seuil spécifique (par exemple un seuil = 0,5). Mais nous pensons que cette mesure peut ne pas être juste ou représentative. Là encore nous avons préféré utiliser l'aire sous la courbe pour classer les algorithmes de suivi (voir Figure 4.11.b).

4. 2.4.4 Résultats

Convergence du filtre particulaire. Avant de commencer les tests sur notre base de données, nous avons étudié la sensibilité du filtre particulaire par rapport à l'initialisation. En effet, lors de l'initialisation, les vecteurs d'état portés par les particules sont créés aléatoirement autour d'un vecteur d'état initial en suivant une distribution normale ($\mathcal{U}[0, \sigma^2]$). Nous voulions vérifier si le filtre particulaire convergeait vers le même suivi quelle que soit l'initialisation. Pour cela nous avons appliqué dix fois le filtre particulaire sur la même séquence. La Figure 4.12 montre les 10 courbes de précision et les 10 courbes de succès de notre test. La Figure 4.13 montre la dispersion des aires sous la courbe de ces

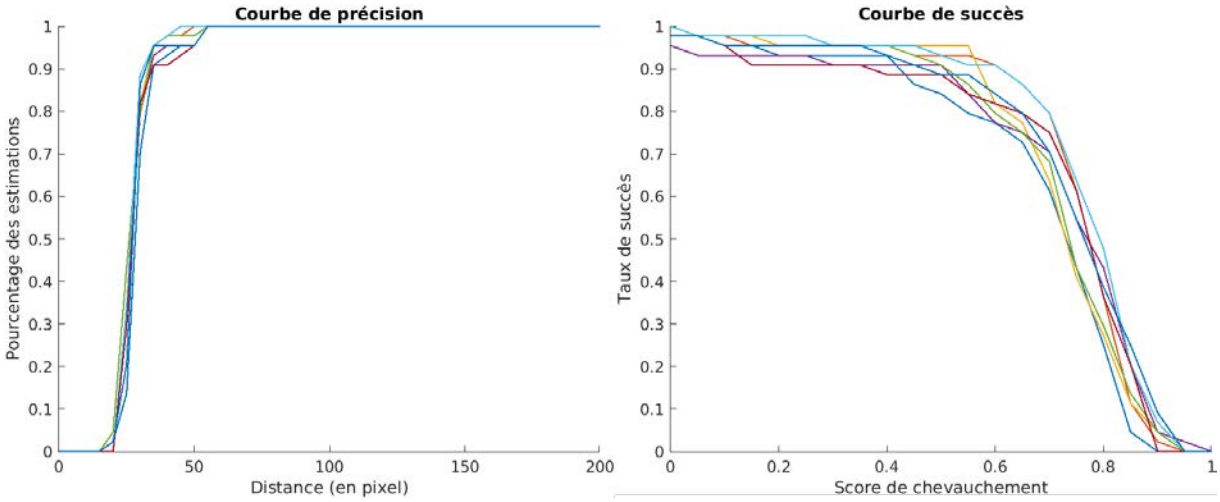


FIGURE 4.12 – Évaluation de l'algorithme de suivi sur la même séquence en calculant la courbe de précision et la courbe de succès pour 10 initialisations différentes.

deux types de courbe sous la forme de boîte à moustache. La médiane des aires sous les courbes de précision est d'environ 0.854 avec des écarts inter-quartiles de l'ordre de 0.0061. Un comportement similaire est observé sur les courbes de succès (médiane de 0.73011 et écarts inter-quartiles de l'ordre de 0.044). Nous constatons que le filtrage particulaire n'est pas sensible aux initialisations. Dorénavant, nous allons tester les différentes variantes de notre algorithme avec une seule réalisation à chaque fois.

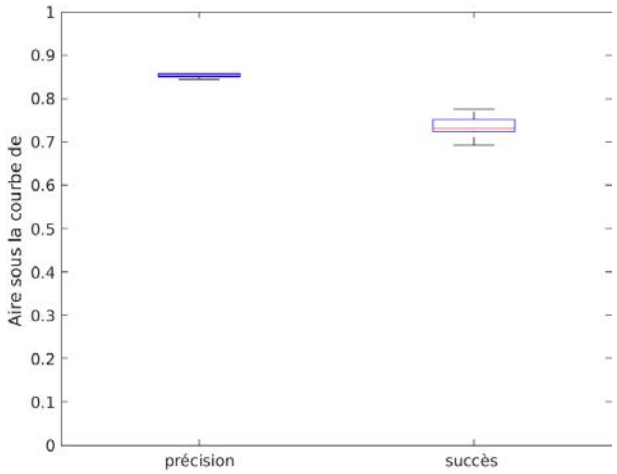


FIGURE 4.13 – Boîte à moustaches des distributions des aires sous les courbes de précision et de succès obtenues sur la même séquence pour 10 initialisations différentes.

Influence du nombre de particules. Pour le filtrage particulaire, le nombre des particules est un paramètre très important pour réussir le suivi de l'objet en mouvement. De manière intuitive nous pensions que plus le nombre des particules est élevé plus l'algorithme est performant. Nous avons testé notre algorithme en modifiant à chaque fois le nombre des particules (de 600 à 1500 particules avec un pas de 100). Les courbes de précision de 10 tests sont visibles dans la Figure 4.14. Cette figure montre une nette amélioration du taux de succès. Cette amélioration est très remarquable entre 900 à 1200 particules. Nous avons fixé $N = 1100$ particules pour les prochains tests puisque son AUC de succès (0.781) est le plus élevé (voir Table 4.1).

Écart-type de la combinaison des coefficients pour la mise à jour de l'état prédit. Un deuxième paramètre de filtrage particulaire, qui influence aussi les résultats de suivi, est la valeur de σ utilisée dans l'équation de mises à jours des pondérations des particules (eq. 4.7). Pour choisir la valeur optimale de ce paramètre, nous avons fait plusieurs tests en modifiant à chaque fois σ de 0.025 à 0.555. La courbe 4.15 illustre les erreurs de localisation du centre de la tête estimé par rapport à celui de la vérité terrain sous forme des courbes de précision. Nous observons que les courbes obtenues avec des valeurs de σ 0.125 et 0.255 sont les plus élevées. Les AUC de ces courbes sont 0.93 et 0.92 respectivement. Nous avons donc fixé σ à 0.125 pour le reste de ce rapport.

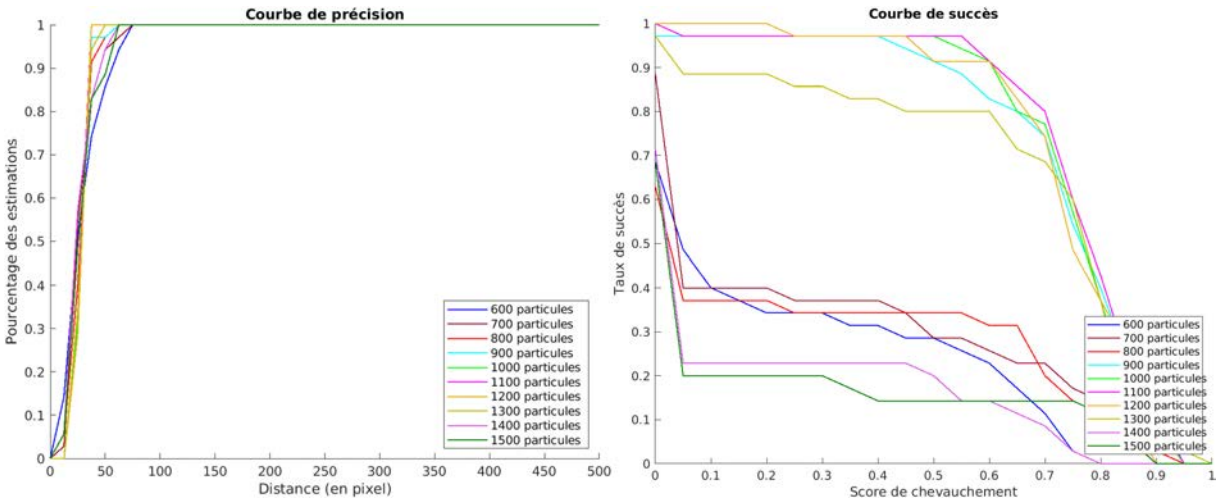


FIGURE 4.14 – Courbes de précision et de succès de dix tests pour évaluer l'impact du nombres de particules sur le modèle de suivi.

Nombre de particules	**AUC de précision**	**AUC de succès**
600	0.931	0.248
700	0.934	0.301
800	0.935	0.284
900	0.933	0.755
1000	0.934	0.772
1100	0.935	0.781
1200	0.933	0.757
1300	0.933	0.694
1400	0.936	0.174
1500	0.933	0.168

TABLE 4.1 – Aire sous les courbes de précision et de succès moyennes de 10 tests différents.

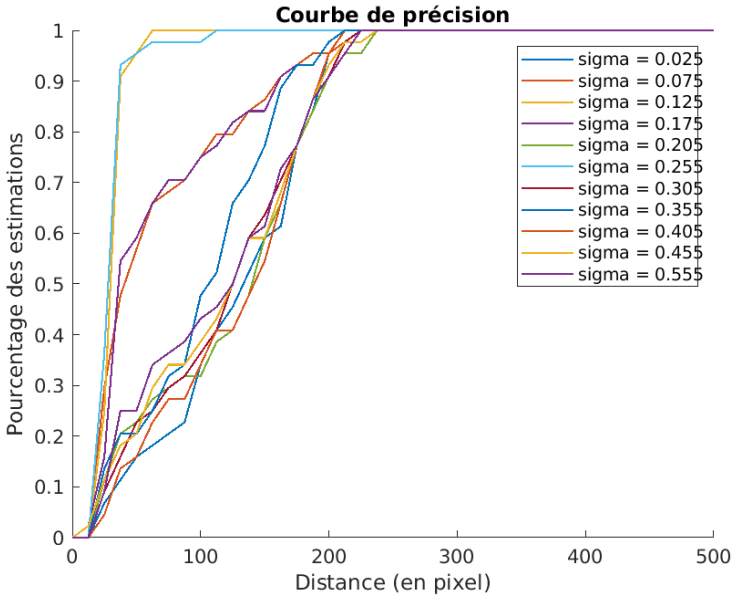


FIGURE 4.15 – Courbes de précision en fonction des différentes valeurs de σ.

Variantes du vecteur d'état. Concernant le vecteur d'état, nous avons combiné les paramètres de l'ellipse qui englobe la tête de plusieurs façons. En effet, nous avons testé quatre variantes différentes de ce vecteur. La Figure 4.16 montre à chaque fois deux exemples visuels du suivi obtenu en utilisant ces variantes. L'image 4.16.*a* représente le résultat du suivi avec un vecteur composé de la seule position de la tête ($x_t = (x_H, y_H)^T$). L'estimation de cette position est erronée comparée à la vérité terrain. Pour compléter la représentation de la personne, nous avons ajouté la taille ($x_t = (x_H, y_H, L)^T$) et l'orientation ($x_t = (x_H, y_H, \theta)^T$) de la tête au vecteur d'état. Les résultats de ces modifications sont visibles sur les images 4.16.b et 4.16.c, respectivement. Concernant l'image 4.16.d, la position estimée résulte de la combinaison de tous ces paramètres ($x_t = (x_H, y_H, L, \theta)^T$).
La Figure 4.17 montre les différentes courbes de précision de chaque variante testée sur trois séquences de 100 images. Ces séquences contiennent différentes activités (marche, accroupi, assis). Les boîtes à moustaches des aires sous la courbe correspondante sont présenté dans la Figure 4.18. Quantitativement, nous constatons que l'utilisation du vecteur d'état le plus complet ($x_t = (x_H, y_H, L, \theta)^T$) semble donner le meilleur recouvrement vu que les AUC sont les plus élevées (0.737 et 0.557 respectivement pour la précision et le succès). Pour la suite, c'est le vecteur d'état le plus complet qui sera retenu.

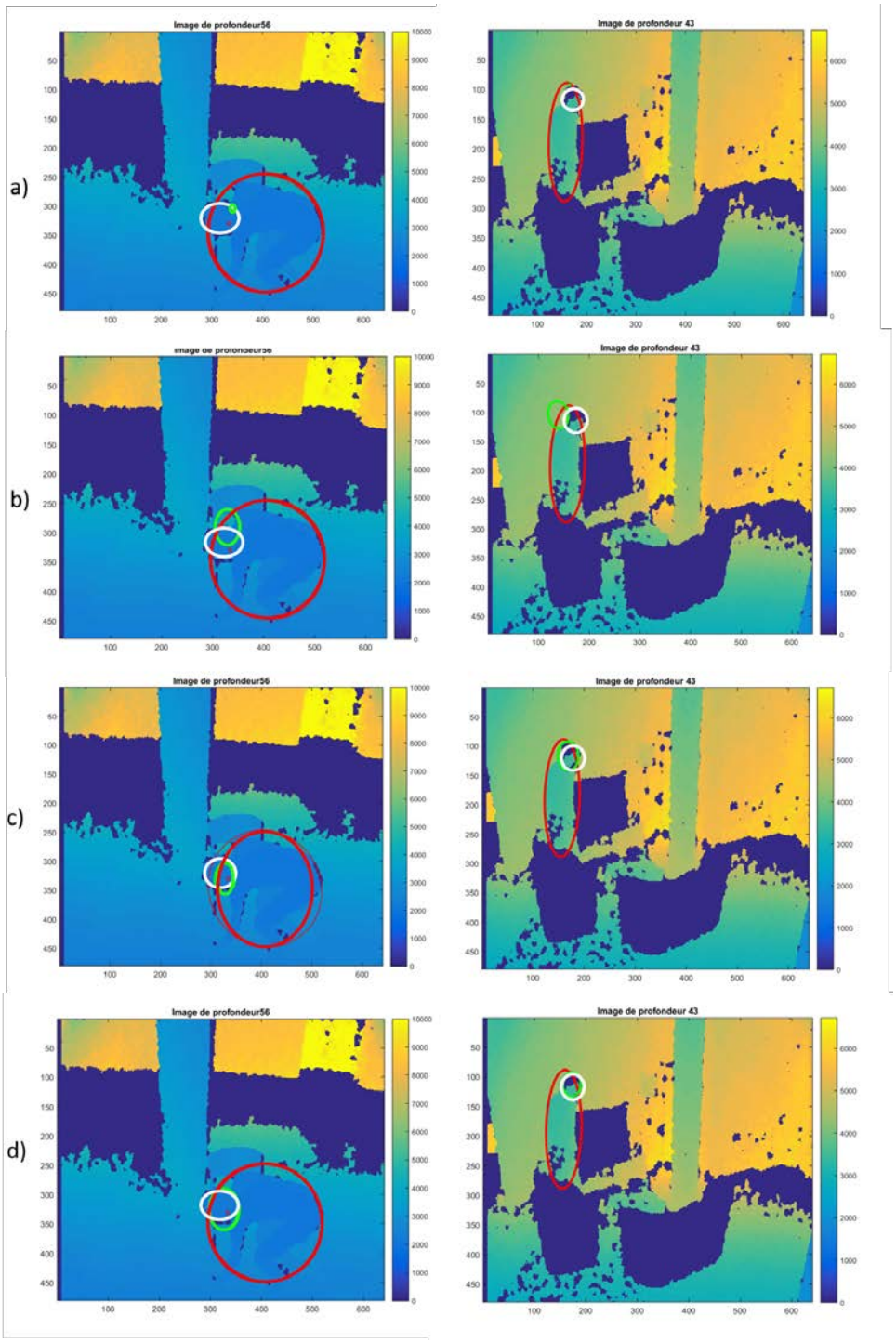


FIGURE 4.16 – Résultats du suivi en définissant le vecteur d'état par a) la position de la tête seule b) la position et la taille de la tête, c) la position et l'orientation de la tête et d) la position, la taille et la taille de la tête. Les résultats du suivi sont en vert, l'ellipse de segmentation de la silhouette est en rouge et l'ellipse GT est en blanc.

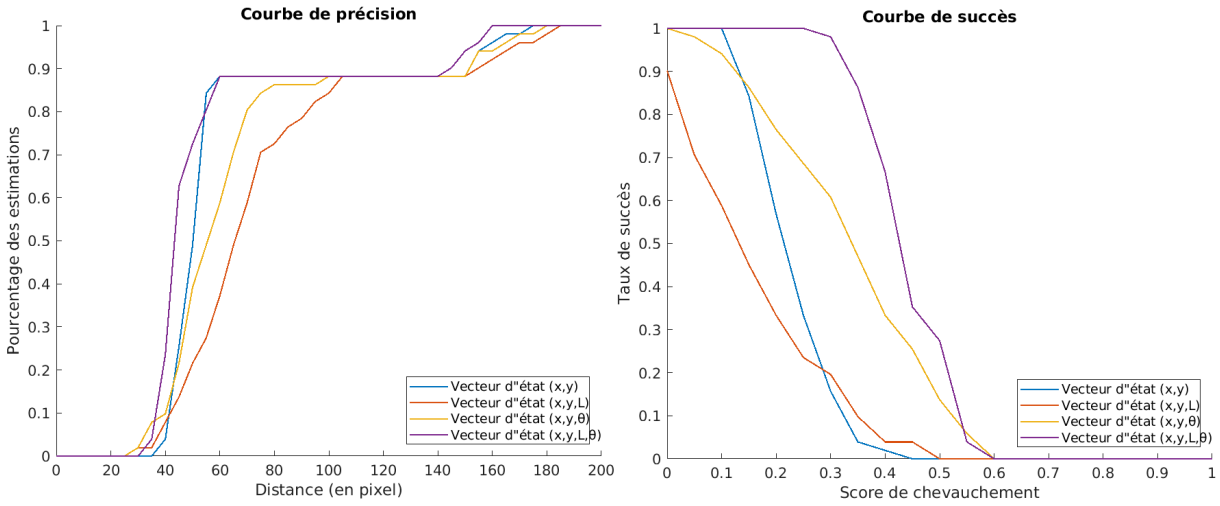


FIGURE 4.17 – Courbes de précision et de succès en fonction des différentes variantes du vecteur d'état

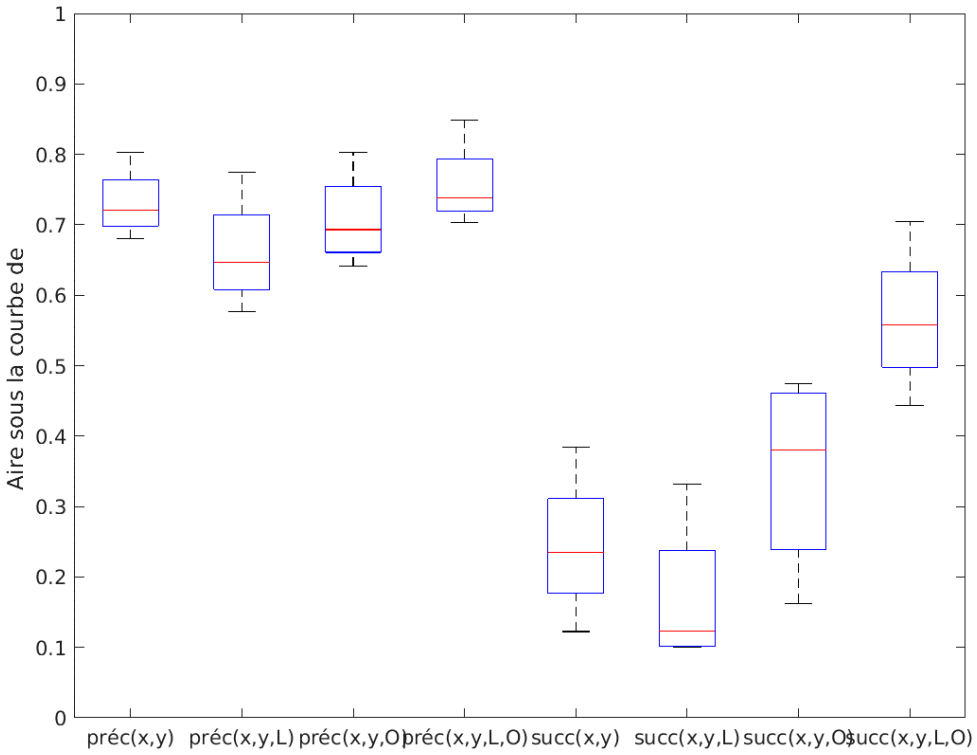


FIGURE 4.18 – Boîtes à moustaches des distributions des aires sous les courbes de précision et de succès de différentes variantes de vecteur d'état : (x, y), (x, y, L), (x, y, θ) et (x, y, L, θ).

Combinaisons des coefficients pour la prédiction. Concernant la phase de prédiction, pour le terme d'observation new_{obs} (eq. 4.7), le choix des coefficients et leur combinaison sont aussi importants. Dans la section 4.2, nous avons détaillé 3 coefficients différents. Nous avons testé trois variantes de new_{obs} correspondants à trois types de combinaisons de ces coefficients :

$$new_{obs_D} = C_D^P(s) \tag{4.14}$$

$$new_{obs_{DG}} = C_D^P(s) + C_G^P(s) \tag{4.15}$$

$$new_{obs_{DF}} = C_D^P(s) + C_F^P(s) \tag{4.16}$$

La Figure 4.19 montre des exemples de résultats visuels de suivi obtenus par les trois combinaisons de coefficients de profondeur. Les images a) et d) de cette Figure contiennent l'estimation de la position de la tête par suivi en ne prenant en compte que de la distance comme observation (new_{obs_D} équations 4.14), les images b) et e) en prenant en compte également le coefficient de gradient ($new_{obs_{DG}}$ éq. 4.15) et les images c) et f) en prenant en compte les coefficients de distance et de fond($new_{obs_{DF}}$ éq. 4.16).
La Figure 4.20 montre les courbes de précision et de succès des différentes variantes de new_{obs} avec la distribution des aires sous la courbe donnée dans la Figure 4.21.

Nous remarquons que la distance seule ne peut pas corriger l'état prédit de la tête (*la mdiane de* $AUC_P = 0.716$ *et* $AUC_S = 0.234$, des courbes de précision et de succès respectivement), puisque la taille de l'ellipse est plus petit que la taille normale de la tête, ce qui explique la forme de la courbe de succès. Cependant l'ajout d'un deuxième coefficient améliore l'estimation de la pose de la tête. Dans notre cas, les coefficients de gradient et de premier plan sont proches, leurs médianes d'$AUC_{prcision}$ sont identiques (0.75305) et d'AUC_{succes} sont 0.456 et 0.373 respectivement. Ceci est confirmé par les aires moyennes sous les courbes de précision et de succès (Table 4.2). La moyenne de l'AUC de coefficient de distance seul est faible par rapport aux autres combinaisons que ce soit pour la courbe de précision ou la courbe de succès.

Toutefois, nous choisissons $new_{obs_{DG}}$ car le coefficient de premier plan de $new_{obs_{DF}}$ augmente le temps d'exécution du modèle d'un facteur 100 (cf. la Table 4.3 qui contient les temps de calcul en utilisant l'une des combinaisons new_{obs} par image).

	AUC moyenne de précision	**AUC moyenne de succès**
Coefficient de distance ($\boldsymbol{new_{obs_D}}$)	0.709	0.145
Coefficient de distance et de premier plan ($\boldsymbol{new_{obs_{DF}}}$)	0.725	0.413
Coefficient de distance et de gradient ($\boldsymbol{new_{obs_{DG}}}$)	0.770	0.463

TABLE 4.2 – Aires moyennes des courbes de précision et de succès de différentes combinaisons de coefficients : D distance, DF distance et premier plan et DG distance et gradient.

	new_{obs_D} (éq.4.14)	$new_{obs_{DG}}$ (éq.4.15)	$new_{obs_{DF}}$ (éq.4.16)
Temps de calcul par image en secondes	0.03	0.2	20

TABLE 4.3 – Temps écoulé pour chaque combinaison de coefficients de profondeur.

Estimation de la tête à partir des particules. Nous avons testé deux façon d'estimer la position de la tête à partir des particules : le modèle AM, la moyenne pondérée des N particules et le modèle MM, qui ne garde que la particule ayant le poids maximum (cf. section 4.2). Les résultats de ces modèles sont présentés respectivement sur les Figures a et b de la Figure 4.22. Nous pouvons constater que le modèle AM fournit des poses plus proches de GT. En effet, le modèle MM choisit la silhouette de la particule qui a le poids le plus élevé, ce qui n'est pas toujours la solution la plus optimale.
Les courbes de précision et de succès sont présentées dans la Figure 4.23. Nous constatons que AM a un meilleur comportement que MM.

4. 2.4.5 Discussion

Dans ce chapitre, nous avons commencé par tester les différentes variantes du filtre particulaire. Dans un premier temps nous avons constaté qu'il fallait 1100 particules pour atteindre un certain plateau de précision. Ensuite nous avons constaté que les composantes du vecteur d'état porté par les particules avaient un impact direct sur les performances de l'algorithme. Dans notre cas, le vecteur d'état le plus complet, composé de la position x_H, y_H, la taille L et l'orientation θ du modèle de tête donnait le meilleur suivi. Concernant le modèle de mise à jour des poids des particules (eq. 4.7) nous avons trouvé d'une part que la variance $\sigma = 0.125$ donnait une bonne précision et d'autre part, que la combinaison des coefficients pour la prédiction n'incluant que la distance et le gradient (eq. 4.15) était un bon compromis entre précision et temps de calcul. Finalement, que le modèle de prédiction de la pose de la tête AM, qui prend en compte la moyenne pondérée de ces N particules, était la plus proche de la vérité terrain, contrairement au modèle MM qui ne gardait que la particule ayant le poids maximum.

Au final, avec cette configuration, nous avons voulu montrer les apports de l'algorithme de suivi par rapport à la simple segmentation de la tête. La Figure 4.24 montre la différence entre le suivi basé uniquement sur la segmentation et l'application d'un algorithme de suivi. Ces résultats prouvent bien que la segmentation est erronée car la taille et la position de la tête ne sont pas au bon endroit. Cette information résulte d'un problème de détection qui engendre une désynchronisation. En effet, la segmentation considère que cette position est relative au point le plus haut de la silhouette. Cependant, l'algorithme de suivi traite l'état actuel de la silhouette ainsi que ses poses dans les images précédentes. Afin de valider ces résultats, nous avons évalué ces modèles en étudiant la courbe de la précision et celle de succès (Figure 4.25). Les résultats de l'évaluation sur une séquence montrent

que l'algorithme de suivi augmente les performances du modèle de segmentation. Pour valider ces résultats, nous avons évalué ces modèles en étudiant la courbe de la précision et celle de succès (Figure 4.25). Les résultats de l'évaluation sur une séquence montrent que l'algorithme de suivi augmente les performances du modèle de segmentation.

4. 3 Conclusion

Dans ce chapitre, nous avons détaillé une approche de suivi basée sur un filtre particulaire et appliquée sur des images de profondeur, pour détecter la tête d'une personne à domicile. Ce chapitre nous a permis de poser les bases de notre système de suivi. La représentation de la personne comporte la position, l'orientation et la taille de l'ellipse qui entoure sa tête.
Nous avons donc opté, pour la présentation de la personne, les trois paramètres pertinents suivants :

— la position du centre de la tête ;
— la taille de la tête ;
— l'orientation de la silhouette.

Ces premières expérimentations nous permettent de conclure que le choix de chaque paramètre influence sur la performance du système. Par ailleurs, dans la mesure où les images de profondeur sont de faible résolution, nous allons exploiter une autre source d'information dans la suite du rapport pour améliorer les résultats obtenus dans ce chapitre.

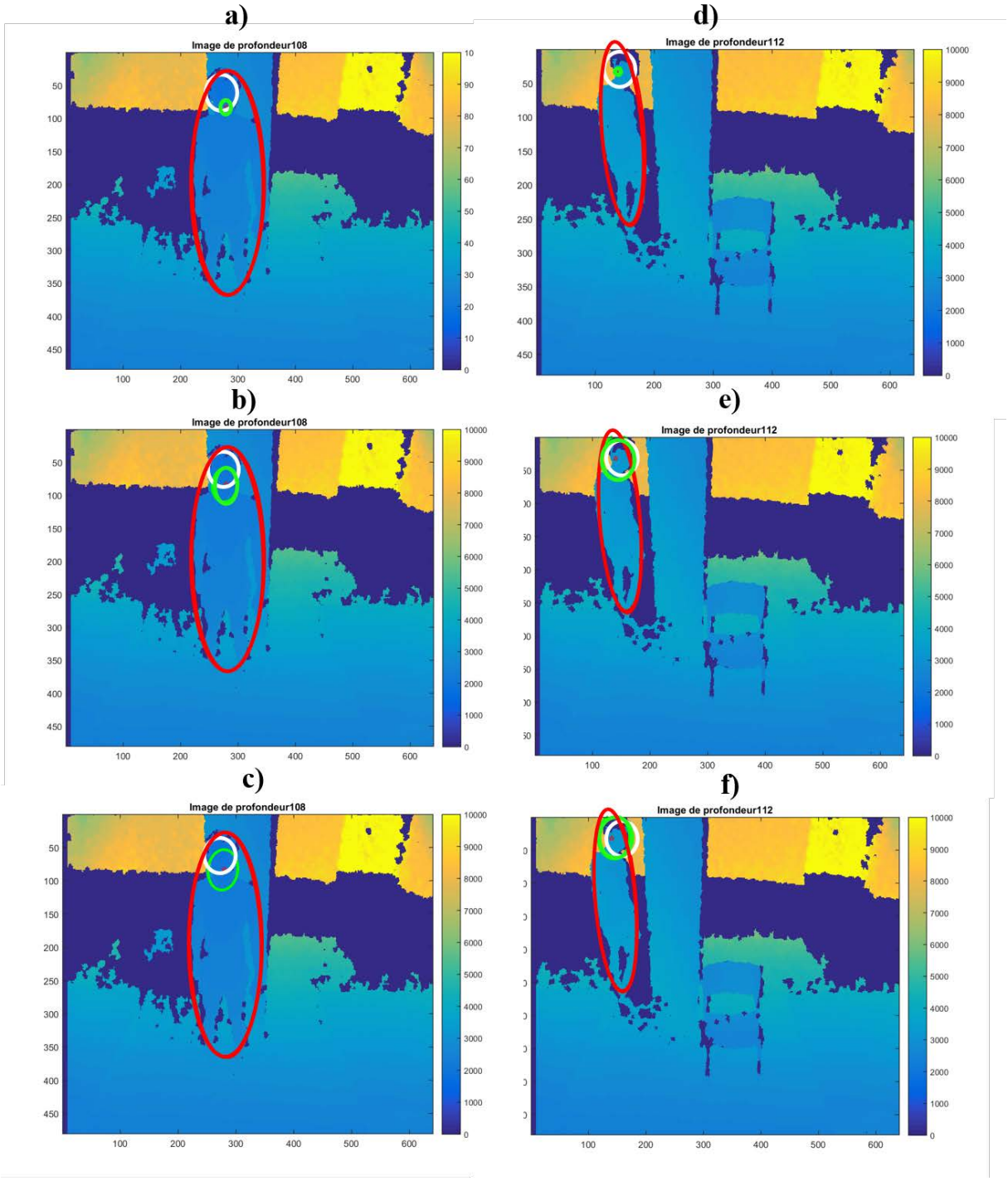


FIGURE 4.19 – Résultats de différentes combinaisons de : a) et d) coefficient de distance new_{obs_D}, b) et e) coefficients de distance et gradient $new_{obs_{DG}}$ et c) et f) coefficients de distance et premier plan $new_{obs_{DF}}$. L'ellipse de la tête estimée par le suivi est en vert et celle de la vérité terrain est en blanc. L'ellipse ajustée sur la silhouette segmentée est en rouge.

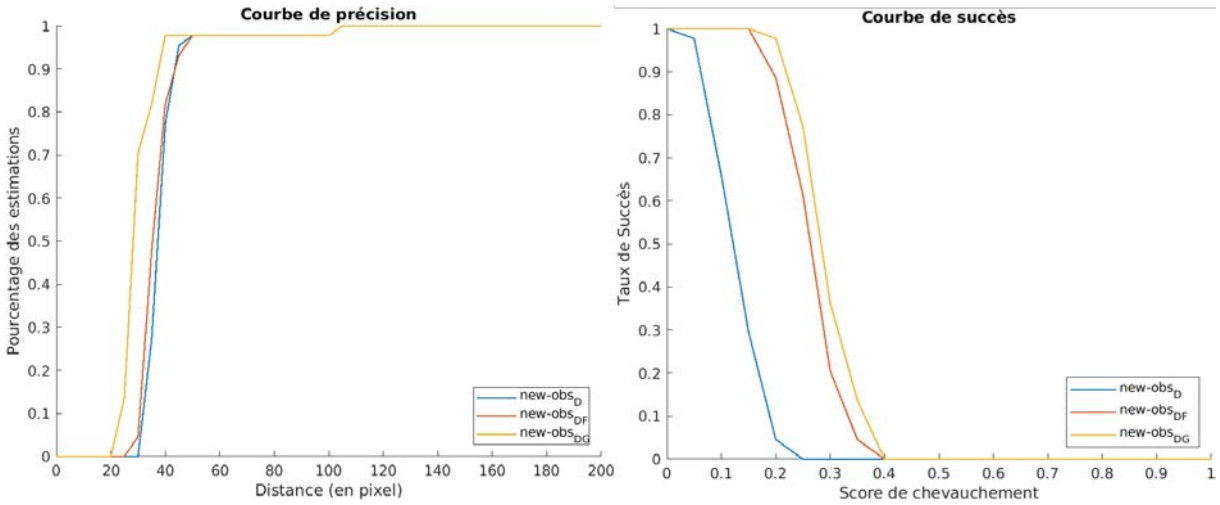


FIGURE 4.20 – Courbe de succès de différentes combinaisons des coefficients de profondeur.

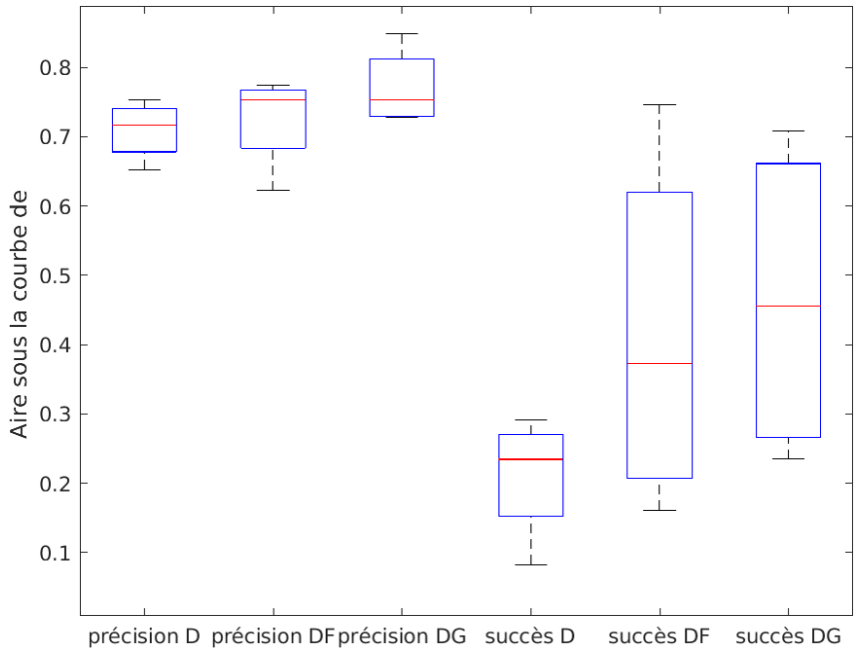


FIGURE 4.21 – Boîtes à moustaches des distributions des aires sous les courbes de précision et de succès de différentes combinaisons de coefficients : D distance, DF distance et avant-plan et DG distance et gradient.

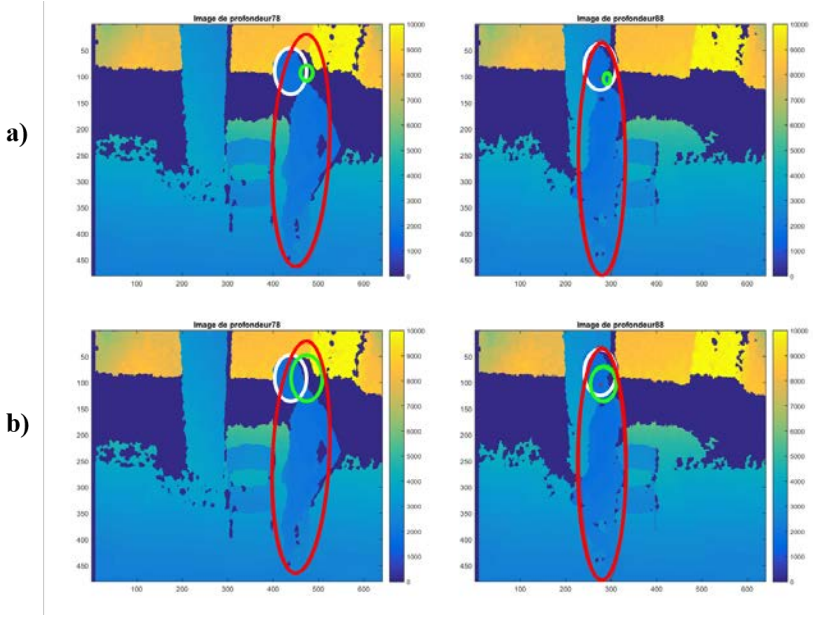


FIGURE 4.22 – Résultats du suivi sur deux images a) modèle *MM* et b) modèle *AM*. Les résultats du suivi sont en vert, l'ellipse de segmentation est rouge et l'ellipse GT est blanche.

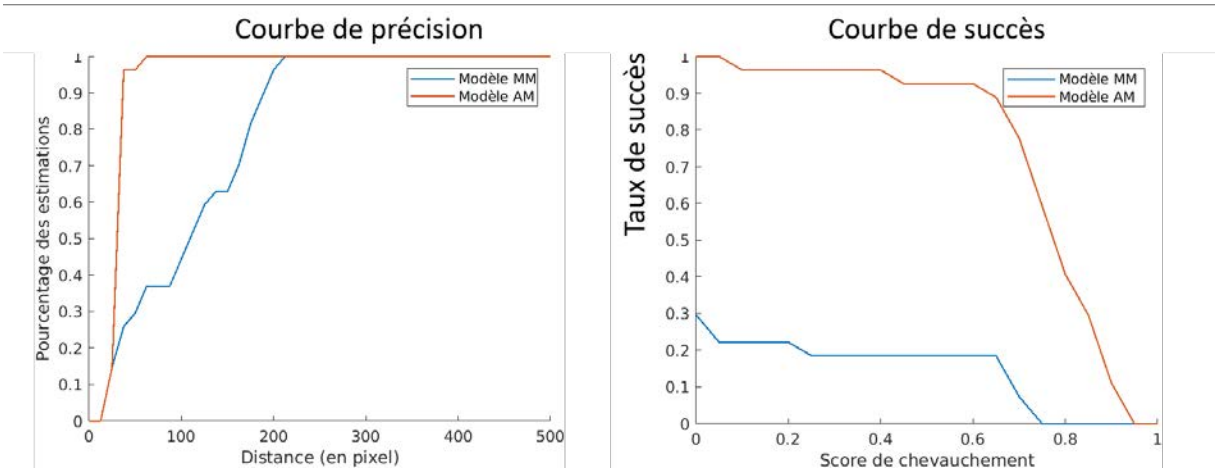


FIGURE 4.23 – Courbe de précision et de succès des modèles *AM* et *MM*.

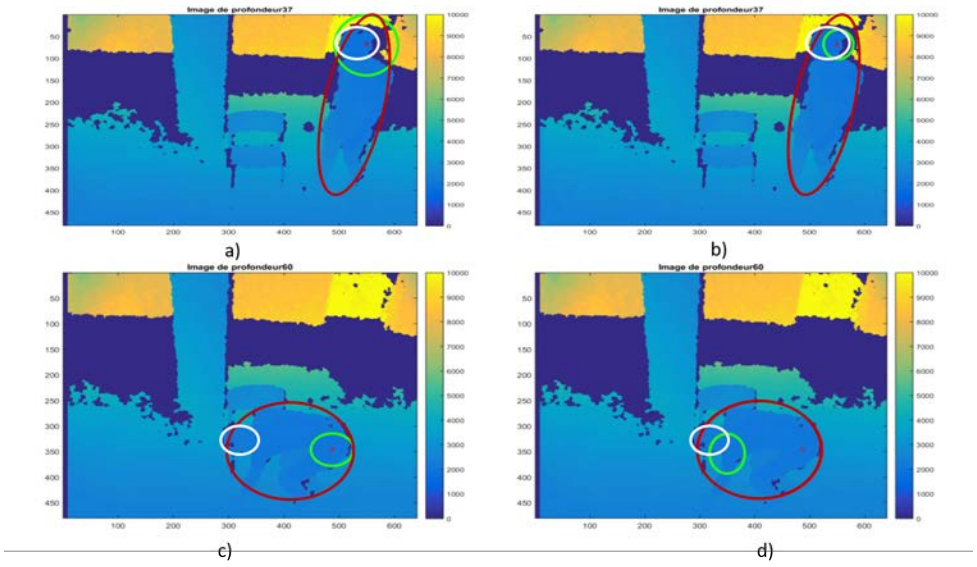


FIGURE 4.24 – Exemple des résultats de a-c) la segmentation seule (ellipse verte) et b-d) de l'algorithme de suivi (ellipse verte) appliqués sur des images de profondeur en détectant l'ellipse de la silhouette (ellipse rouge) en définissant la vérité terrain (ellipse blanche).

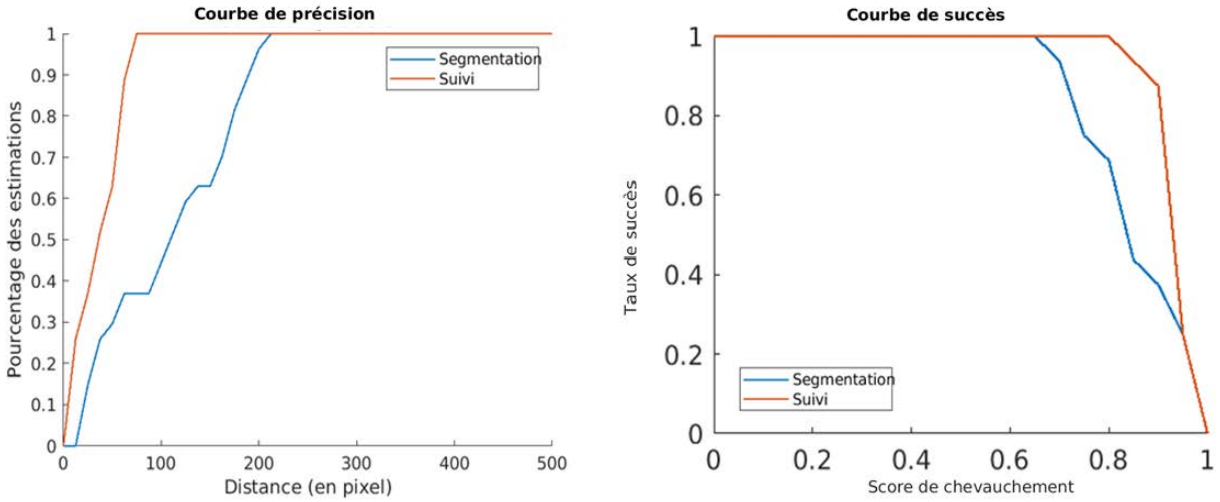


FIGURE 4.25 – Courbes de précision et de succès des résultats de la segmentation (courbe bleue) et de l'algorithme de suivi (courbe rouge) appliqués sur des images de profondeur.

Chapitre 5

Fusion des informations thermiques et de profondeur

Dans le chapitre précédent, nous avons détaillé l'algorithme de suivi de la tête de la personne par filtrage particulaire en utilisant des images de profondeur uniquement. Dans ce chapitre, nous enrichissons cet algorithme par une deuxième information issue d'un capteur thermique pour améliorer les résultats obtenus précédemment.

Introduction

Le choix de l'image thermique s'appuie sur la préservation de l'anonymat de la personne, la facilité d'extraire la silhouette, le faible coût et la taille du capteur. Le capteur thermique permettra surtout de discriminer d'éventuels objets froids d'objets chauds (individu). Le FLIR Lepton 2.5, que nous avons choisi pour notre solution, peut s'installer discrètement dans la chambre.

Dans le chapitre précédent, nous avons choisi de suivre la tête sur les images de profondeur. En fait, ce choix s'applique également sur nos images thermiques car, dans la plupart des cas, la tête représente la zone la plus chaude de la silhouette sur l'image thermique, ce qui justifie notre choix de cette région d'intérêt pour les images thermiques.

Dans ce chapitre, nous appliquons notre algorithme de suivi sur les images thermiques et de profondeur en fusionnant ces deux informations lors de l'étape de mise à jour des poids de particules. Ensuite, nous comparons ces résultats avec ceux obtenus à partir des images de profondeur. Cette méthode n'est toutefois basée que sur des informations de position, taille et orientation de la tête. Or le suivi est un processus dynamique. La vitesse de déplacement devrait être également prise en compte. Nous proposons donc quelques améliorations à notre algorithme de fusion en nous appuyant sur un modèle adaptatif basé sur la position, et aussi sur la vitesse de déplacement de la tête. Pour ce faire, nous avons divisé ce chapitre en deux sections principales : une pour le modèle statique et une

autre pour le modèle dynamique.
La principale contribution de ce chapitre est la fusion des informations de profondeur et thermiques pour améliorer le suivi visuel. À chaque instant, une nouvelle combinaison de coefficients pour chaque particule est établie, basée sur une pondération adaptative. Les résultats montrent que l'algorithme de suivi peut traiter les cas de mouvement rapide (comme le cas de la chute), d'occultations partielles et de variation d'échelle.

5. 1 Modèle de suivi en fusionnant les informations thermiques et de profondeur

Notre modèle de suivi de la tête est basé sur la fusion de l'information thermique avec celle de profondeur comme illustré Figure 5.1. Ce modèle est une nouvelle version du modèle décrit dans le chapitre précédent (c.f Figure 4.1). Les modifications faites sur ce modèle sont principalement situées au niveau de données d'entrée de l'algorithme et de l'étape de suivi.

5. 1.1 Recalage des images

Pour mettre en correspondance tout pixel de l'image de profondeur avec celui de l'image thermique, nous appliquons, dans la première phase de l'algorithme, un recalage entre l'image thermique et l'image de profondeur. Ce recalage consiste à appliquer sur les images les paramètres de transformation estimés lors de la calibration (éq. 2.1-2.5) détaillée dans la section 2.3. Toutefois, les deux images ne sont pas de même résolution (Figure 5.2). Nous rappelons qu'un pixel de l'image thermique correspond à une région de 8×8 pixels sur l'image de profondeur.
La fusion se fait alors de la manière suivante : 1) le filtre particulaire est appliqué sur la résolution la plus grande (celle de l'image de profondeur), 2) l'information de profondeur est donc accessible directement, 3) la position d'un pixel dans l'image en profondeur est projetée sur l'image thermique et l'information thermique correspondante à cette position est donnée par la valeur portée par le plus proche voisin de la position projetée. Le choix d'utiliser le plus proche voisin plutôt qu'une interpolation (bilinéaire ou polynomiale d'ordre plus élevé) est justifié par des raisons de rapidité de temps de calculs.
Pour évaluer la performance de cette étape, nous avons annoté manuellement des points remarquables sur des paires d'images profondeur/thermique d'une séquence (cf Figure

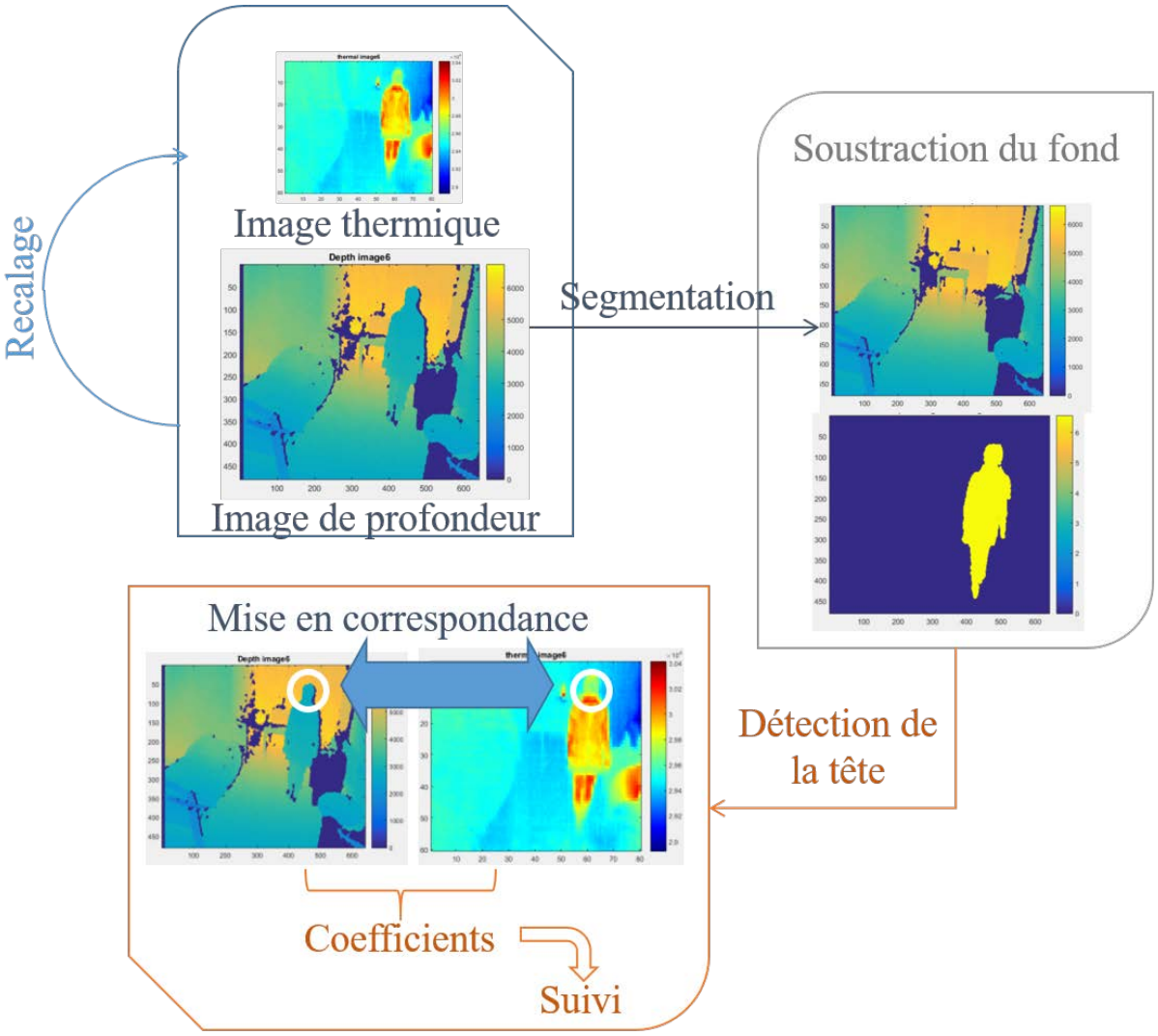


FIGURE 5.1 – Processus de suivi d'un objet en mouvement.

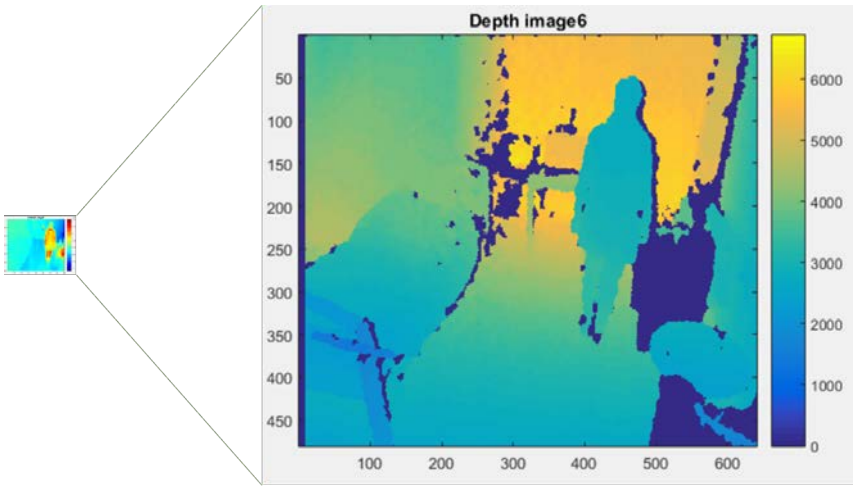


FIGURE 5.2 – Différences de résolution entre l'image thermique (à gauche) et l'image de profondeur (à droite).

5.3). Cette annotation devait être très précise, surtout dans l'image thermique, car une erreur d'un pixel dans cette image correspond à une erreur de 64 pixels dans l'image de profondeur. La position de chaque pixel annoté dans l'image de profondeur est ensuite projetée dans l'image thermique. Ceci nous permet de mesurer une erreur de mise en correspondance. Il s'agit de la distance entre le plus proche voisin de la position projetée et le pixel annoté dans l'image thermique. L'erreur de mise en correspondance moyenne est égale à $0,147540$ pixels.

5. 1.2 Suivi de la tête de la personne

Les étapes du suivi sont identiques à celles décrites dans le chapitre 4 : 1) segmentation de la tête et 2) suivi par filtre particulaire.
La segmentation de la tête est appliquée sur la seule image de profondeur puisqu'elle a la résolution la plus grande. Nous avons donc appliqué la procédure décrite dans la section 4.1.
Le suivi par filtre particulaire est ensuite appliqué sur l'image de profondeur afin de bénéficier de sa résolution. Les principes de l'initialisation des particules, du ré-échantillonnage des particules, de la prédiction du vecteur d'états en fonction de l'équation de transition, de la mise à jour des poids et de l'estimation de la position de la tête sont identiques à ceux décrits dans la section 4.2. Les différences résident dans la définition du vecteur d'état (intégration du vecteur vitesse dans certains cas) et dans l'ajout de l'information

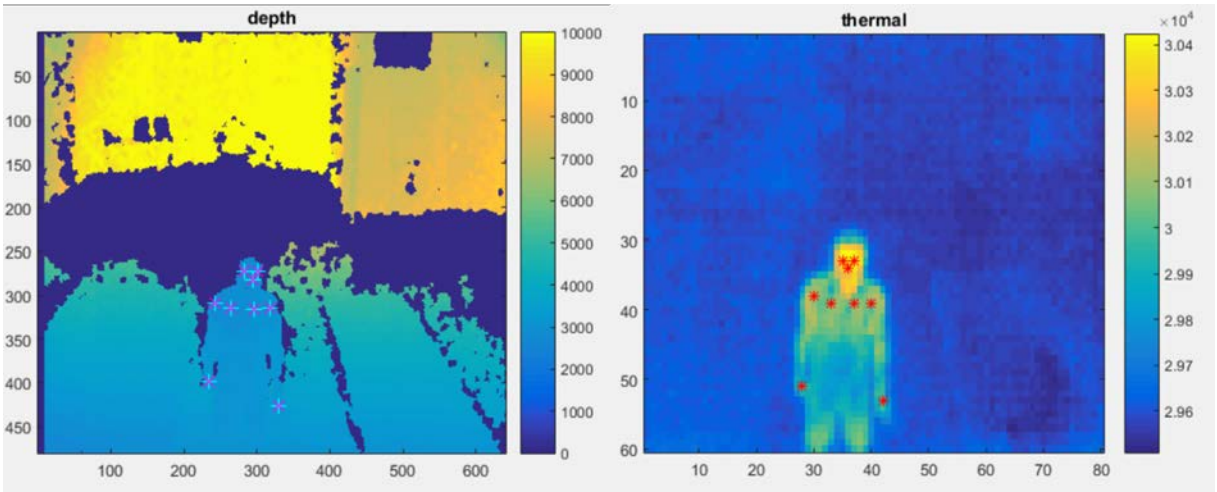


FIGURE 5.3 – Points remarquables annotés manuellement sur des paires d'images profondeur/thermique d'une séquence. Les points correspondant sont identifiés par les points rouges.

thermique à l'étape de mise à jour des particules. En effet, l'observation new_{obs} va intégrer des coefficients issus de l'image thermique en plus de ceux issus de l'image de profondeur.

5. 1.2.1 Coefficients thermiques

Nous avons détaillé dans la section 4.2.3 les coefficients de profondeur C_F^P, C_G^P et C_D^P. Dans cette section, nous avons choisi d'y ajouter deux autres coefficients issus de l'image thermique : C_G^T (coefficient de gradient thermique) et C_T (coefficient de température).

— **Coefficient de gradient thermique C_G^T.** Il s'agit de calculer la somme du gradient orthogonal au périmètre de l'ellipse de la particule sur l'image thermique. Parmi ces valeurs, nous ne gardons que les gradients négatifs (de l'intérieur vers l'extérieur de l'ellipse) puisque la tête a une température plus élevée que l'environnement à l'extérieur :

$$C_G^T = \frac{1}{N}\sum_{k=1}^{N}(n(i).g_s(i))^- \tag{5.1}$$

$$(.)^- = |min(.,0)| \tag{5.2}$$

où $n(i)$ le vecteur orthogonal à l'ellipse et $g_s(i)$ le vecteur du gradient au pixel i.

— **Coefficient de température C_T.** C'est la valeur d'intensité du centre de l'ellipse

de la particule sur l'image thermique. Comme cette valeur n'est pas stable par rapport aux voisinages, nous avons alors affecté à ce coefficient la moyenne des valeurs des pixels à l'intérieur de cette ellipse.

Dans le chapitre précédent, nous avions comparé les trois coefficients de profondeur (C_G^D, C_F^D et C_D^D) et nous avons décidé de ne garder que les deux premiers puisque le coefficient de premier plan (C_F^D) augmentait le temps de calcul du processus sans améliorer les performances du modèle de suivi.
Dans la suite de ce chapitre, nous fusionnons ces deux coefficients thermiques avec les deux coefficients de profondeur retenus afin d'étudier l'impact de chacun de ces coefficients et garder la combinaison la plus pertinente pour notre contexte. Nous proposons deux façons pour fusionner ces coefficients : i) la première, que nous appelons méthode *statique*, consiste à garder la même combinaison (les mêmes pondérations entre coefficients) pour toutes les séquences et ii) la seconde, que nous appelons méthode *dynamique*, consiste à estimer et modifier les pondérations de manière adaptative durant le déroulement de la séquence. Le terme de dynamique est d'autant plus vrai que nous avons également décidé d'incorporer la vitesse de déplacement de la tête dans le vecteur d'état des particules dans cette seconde méthode.

5. 2 Méthode statique pour fusionner les informations thermiques et de profondeur

5. 2.1 Fusion des informations

Dans cette partie, nous avons quatre coefficients pour chaque particule : C_D^P, C_G^P, C_G^T et C_T. Afin d'évaluer l'importance de chaque coefficient, nous avons affecté un facteur d'importance (FI) à chacun d'entre eux. La méthode de fusion statique combine ainsi les quatre coefficients comme suit :

$$new_{obs}(n) = \alpha C_D^P(n) + \beta C_G^P(n) + \gamma C_G^T(n) + \lambda C_T(n) \tag{5.3}$$

avec α, β, γ, $\lambda \in [0, 1]$ les facteurs d'importance (FI) avec $\sum(\text{FI}) = 1$. Ces facteurs d'importance vont nous permettre d'ajuster l'importance de chaque coefficient pour la remise à jour des poids des particules.

Dans un premier temps, nous avons fait une petite étude prospective afin d'étudier

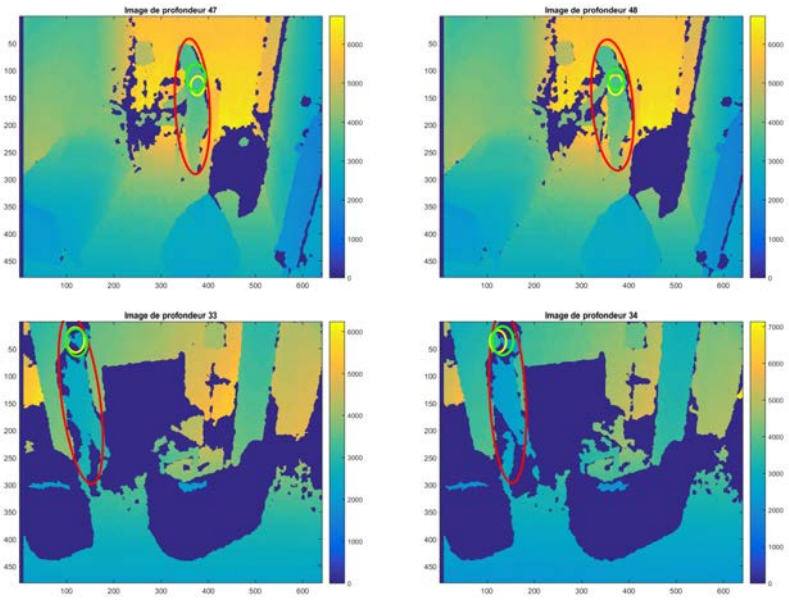


FIGURE 5.4 – Exemples de suivi utilisant soit le coefficient de température seul (ellipse jaune) soit celui de gradient thermique seul (ellipse verte).

le comportement de chaque coefficient. Visuellement nous avons constaté que d'une part le coefficient de distance C_D^P doit toujours être pris en compte car il est discriminant et d'autre part que les coefficients de gradient de température C_G^T et de température C_T avaient un comportement assez similaire pour l'estimation de la position de la tête. Intuitivement, nous pensons que le coefficient de gradient thermique contient également l'information de température puisque cette dernière est utilisée pour son calcul. La Figure 5.4 présente deux exemples de suivi en utilisant ces deux coefficients séparément. Nous pouvons y remarquer que les modèles basés sur ces deux coefficients estiment la position de la tête d'une façon similaire. Nous pouvons alors éliminer l'un de ces coefficients à chaque test pour réduire le temps d'exécution.

Nous avons ensuite fait une étude plus formelle pour estimer la combinaison qui donne les meilleurs résultats en termes de suivi de la tête. Pour cela nous avons proposé plusieurs combinaisons de coefficients en donnant plus d'importance à certains coefficients ou en supprimant d'autres (C_G^T ou C_T). Les 11 combinaisons que nous avons choisies sont détaillées dans la Table 5.1.

Nous avons repris la même méthodologie d'évaluation que pour le chapitre 4 : application du suivi de la tête sur les 10 séquences où la vérité terrain, (la position de la

Test	α	β	γ	λ	**Plus d'impact pour**
C_1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	tous
C_2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	les trois premiers
C_3	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	les trois derniers
C_4	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	α
C_5	$\frac{1}{2}$	$\frac{1}{4}$	0	$\frac{1}{4}$	α
C_6	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	β
C_7	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	0	β
C_8	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	γ
C_9	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	0	γ
C_{10}	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$	λ
C_{11}	$\frac{3}{8}$	$\frac{3}{8}$	0	$\frac{1}{4}$	λ

TABLE 5.1 – Différentes combinaisons de coefficients thermiques et de profondeur.

	Courbe de précision	**Courbe de succès**
Segmentation	0.372	0.055
Profondeur seule	0.681	0.223
Fusion C1	0.849	0.701

TABLE 5.2 – Aires moyennes sous les courbes de précision et de succès moyennes de la segmentation seule, de la profondeur seule et de la fusion profondeur/thermique (modèle C1).

tête) était annotée manuellement, validation qualitative visuelle des suivis et validation et comparaison quantitatives à l'aide des courbes de précision et de succès et de leurs aires sous la courbe.

5. 2.2 Résultats

Nous présentons, dans cette section, les résultats de différentes combinaisons détaillées dans la Table 5.1. Nous avons évalué nos différentes méthodes sur une base de données composée de dix scénarios différents, chacun comportant 100 couples d'images. Les caméras sont fixées au plafond de chaque pièce. Ces séquences contiennent des activités de la vie quotidienne de la personne (s'assoir, allonger, boire, regarder la télé, marcher, ramasser un objet par terre ..). L'objectif de ces tests est d'évaluer la performance de l'algorithme de fusion en comparant l'impact de chaque coefficient à l'aide de leurs facteurs d'importance (FI).

Tout d'abord, nous avons comparé les résultats de la première combinaison (C_1) de la Table 5.1 avec le modèle basé sur les seuls coefficients de profondeur (c.f section 2.4.4). Comme l'illustre la Figure 5.5, nous constatons une nette amélioration des résultats en ajoutant l'information thermique à l'équation 4.16. Ceci est confirmé par les courbes de précision et de succès avec des AUC nettement supérieures pour la combinaison prenant en compte les informations de l'image thermique en plus de l'information de profondeur (Figure 5.6 et Table 5.2).

Ensuite, nous avons comparé les 11 essais de combinaison de coefficients ($C_1, .., C_{11}$) définis dans la Table 5.1. La Figure 5.7 montre l'estimation de la position de la tête sur la même image en utilisant les 11 combinaisons. Toutefois, nous ne pouvons pas confirmer la fiabilité des ces combinaisons sur une seule image. Pour cette raison nous avons évalué ces résultats en utilisant les deux mesures quantitatives détaillées Section 4.2.4.3 (Figure

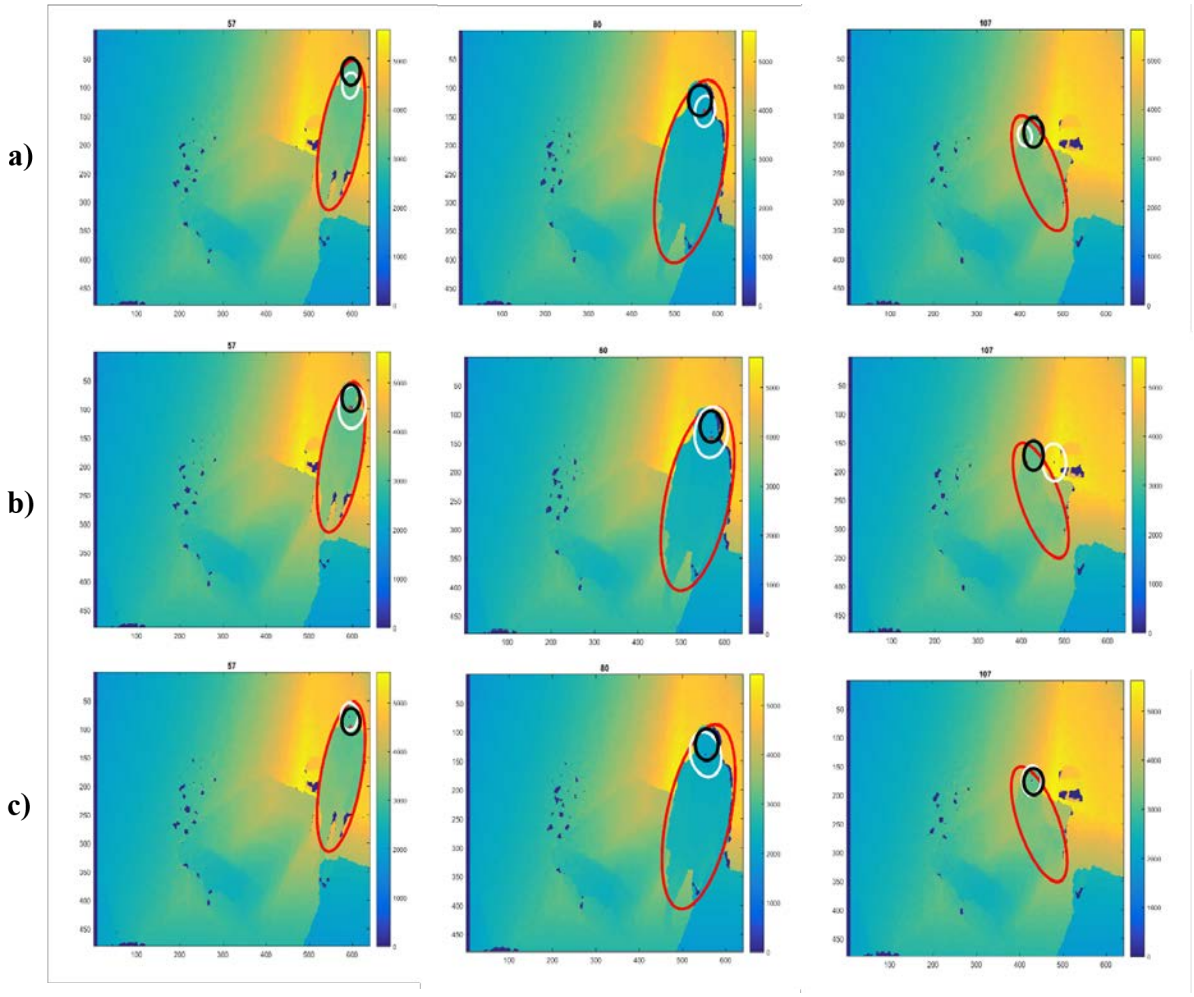


FIGURE 5.5 – Exemple de résultats du suivi sur trois images différentes a) segmentation seule, b) de la profondeur seule, c) premier modèle de fusion (C1). Le résultat du suivi est en blanc, l'ellipse de la silhouette est en rouge et l'ellipse de vérité terrain est en noir.

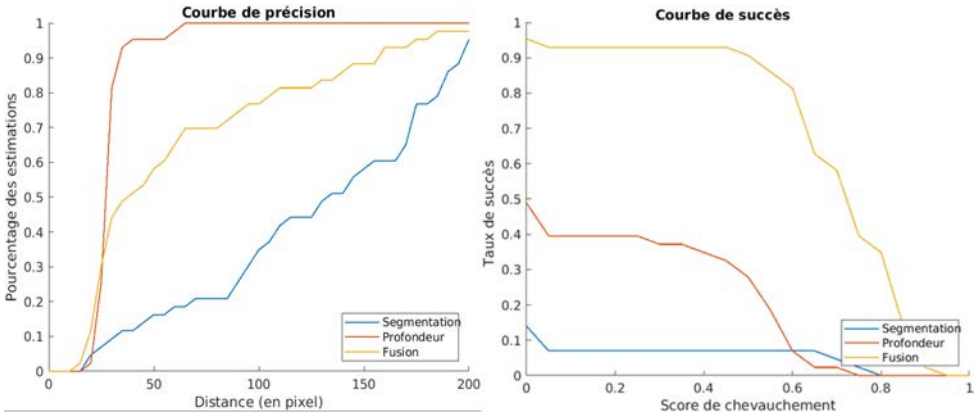


FIGURE 5.6 – Courbes moyennes de précision et de succès pour : a) la segmentation seule (bleu), b) le modèle de profondeur seul (rouge) et c) le premier modèle de fusion (jaune) : courbe de précision, courbe de succès

5.8), ainsi que leurs aires sous la courbe des moyennes (Table 5.3).

Nous pouvons constater que ces mesures quantitatives ne permettent pas de dégager clairement une combinaison par rapport aux autres. Nous avons donc fait une analyse visuelle du comportement du suivi tout le long des séquences. Nous avons constaté que l'importance de chaque coefficient peut changer d'une image à une autre. Par exemple, à certains moments, l'information thermique semble être plus informative que celle de profondeur et inversement. De plus, les observations de profondeur peuvent ne pas être pertinentes lorsque la silhouette de la personne est proche d'un meuble qui a été déplacé après la création de la carte de référence. De même, l'observation thermique peut être inefficace si la personne marche à côté d'un chauffage. Il serait donc préférable de pouvoir adapter la combinaison des coefficients en fonction du contenu des images. C'est le propos de la section suivante sur le modèle dynamique.

5. 3 Modèle de fusion dynamique

Lors de l'analyse des résultats du modèle statique nous avons constaté deux limitations de cette approche :

1) la méthode de suivi est parfois lente quand il s'agit d'un mouvement brusque. Dans certains cas, une personne qui s'accroupit ou qui chute, l'algorithme récupère la position exacte de la tête après cinq images, comme le montre Figure 5.9 (ce qui correspond à un peu moins d'une seconde sachant que nous avons une fréquence d'images de 8 Hz).

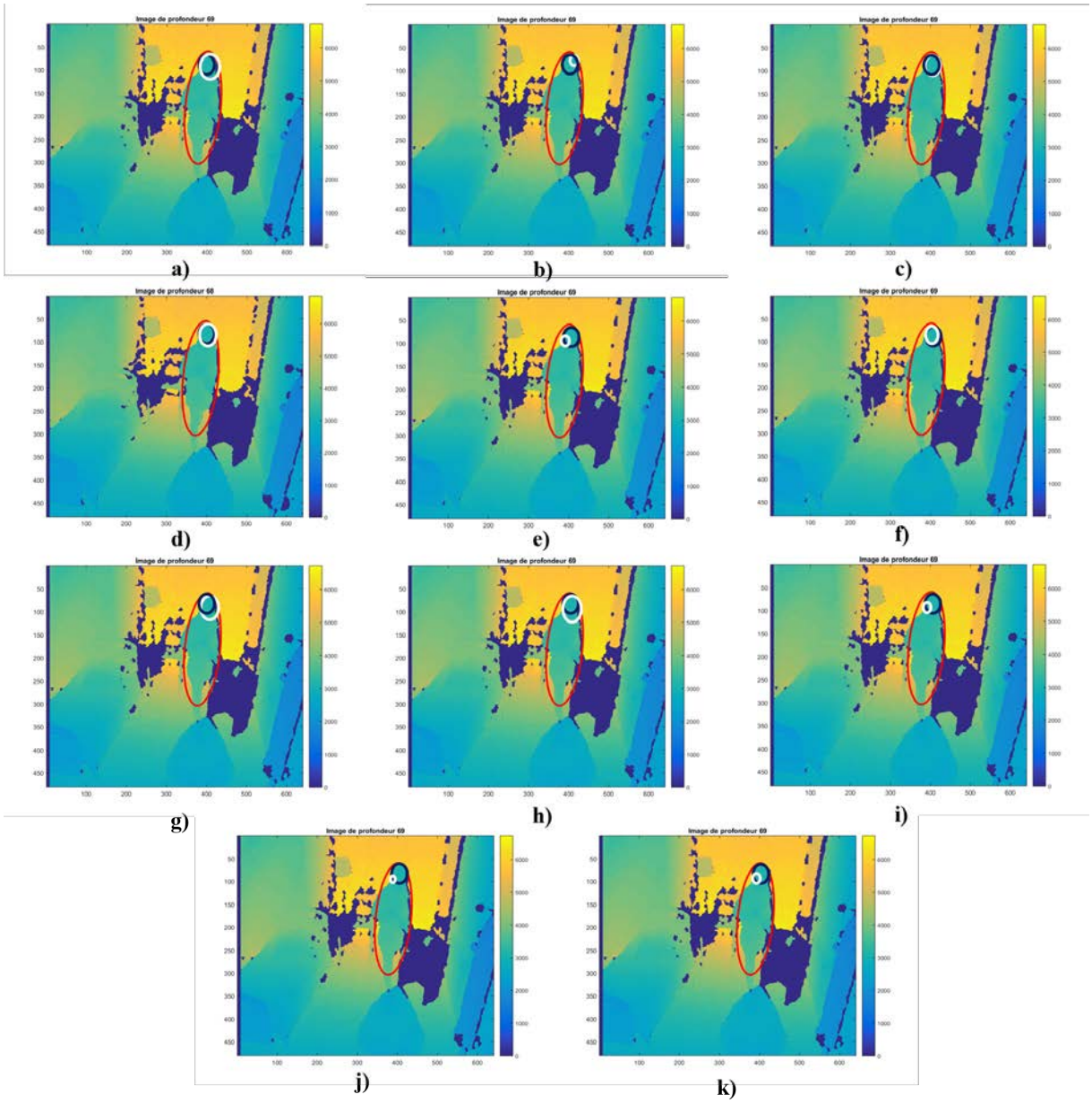


FIGURE 5.7 – Exemple de résultats de suivi sur la même image en utilisant les combinaisons : a) C1, b) C2, c) C3, d) C4, e) C5, f) C6, g) C7, h) C8, i) C9, j) C10, et k) C11. Le résultat du suivi est en blanc, l'ellipse de la silhouette est en rouge et l'ellipse de vérité terrain est en noir.

	Courbe de précision	**Courbe de succès**
C1	0.823	0.619
C2	0.857	0.679
C3	0.857	0.679
C4	0.824	0.429
C5	0.852	0.468
C6	0.843	0.679
C7	0.838	0.693
C8	0.817	0.543
C9	0.822	0.479
C10	0.837	0.645
C11	0.837	0.679

TABLE 5.3 – Aires moyennes sous les courbes de précision et de succès de 11 combinaisons de coefficients thermiques et de profondeur.

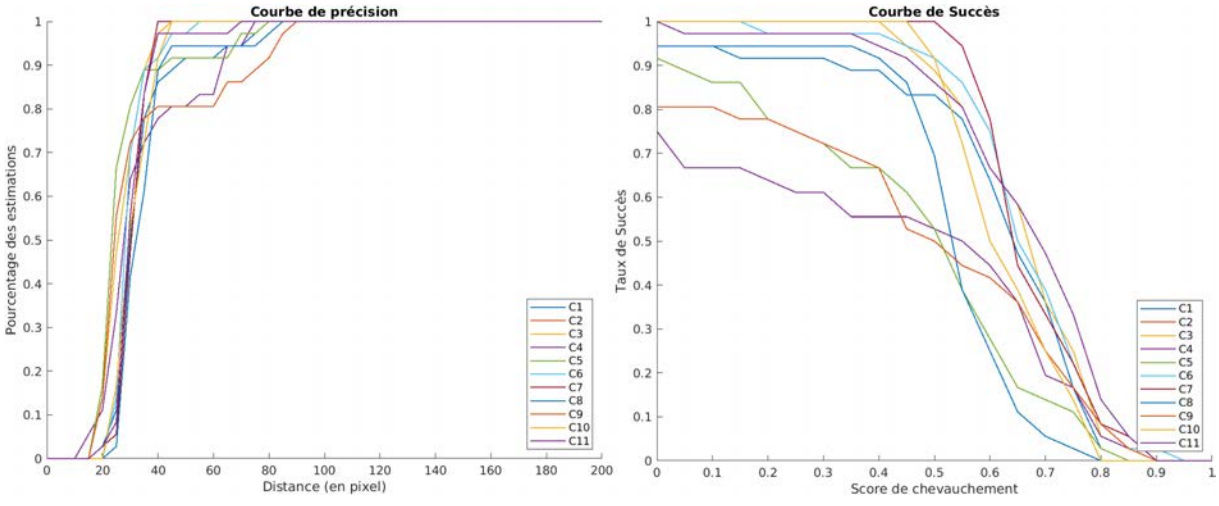


FIGURE 5.8 – Courbes moyennes de précision et de succès de l'algorithme de suivi en testant les 11 combinaisons différentes de coefficients.

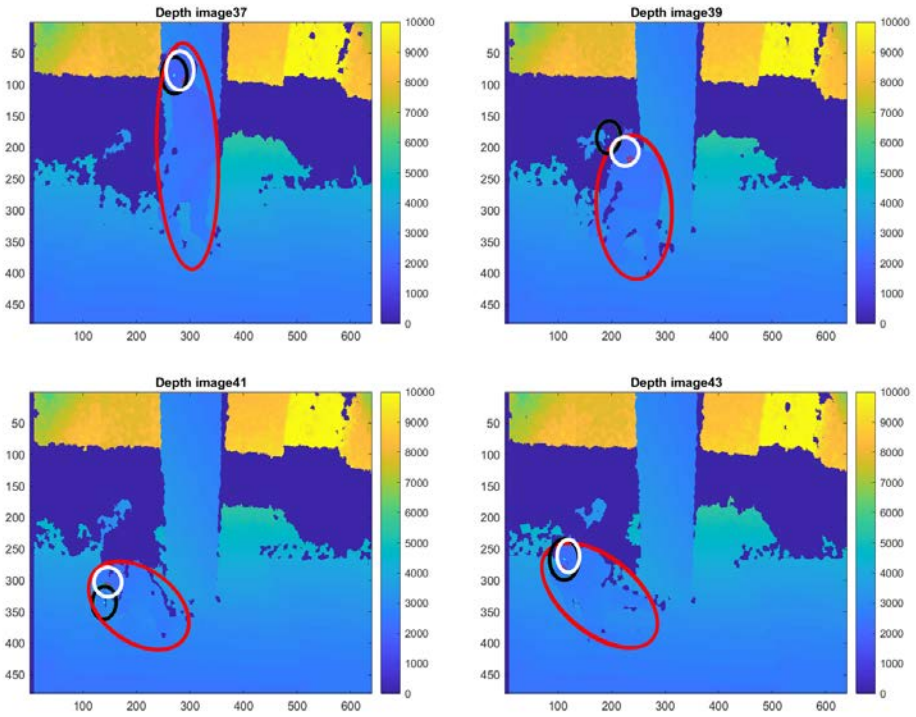


FIGURE 5.9 – Suivi de la tête de la personne sans information de vitesse. Le résultat du suivi est en noir, l'ellipse de la silhouette est en rouge et l'ellipse de vérité terrain est en blanc.

2) le fait que l'information la plus pertinente pour le suivi varie en fonction du contexte de l'image (voir chapitre suivant).

Afin de répondre à la première limitation, nous avons décidé d'intégrer la vitesse de la tête ($\dot{x}_h$,$\dot{y}_h$) dans le vecteur d'état. La représentation finale de vecteur d'état est :

$$x_t = (x_h, y_h, \dot{x}_h, \dot{y}_h, L, \theta) \tag{5.4}$$

Pour répondre à la seconde limitation, nous avons décidé d'ajuster les facteurs d'importance d'une manière dynamique et d'adapter la combinaison des coefficients selon l'importance de chaque coefficient à chaque instant en fonction du contenu. Cette nouvelle méthode de fusion est schématisée dans la Figure 5.10.

Pour faire simple, la combinaison des coefficients est modifiée à chaque image selon

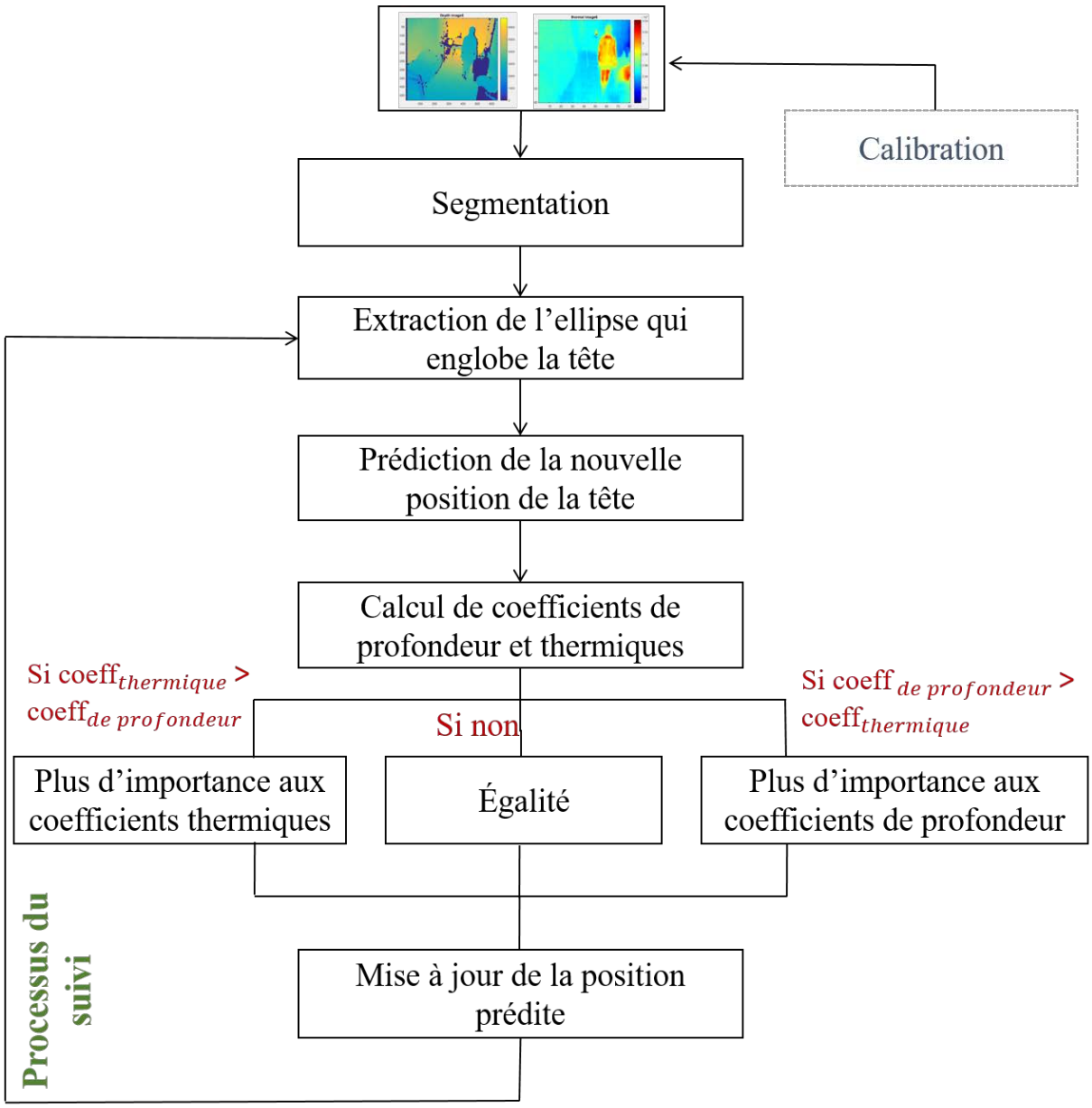


FIGURE 5.10 – Méthode de fusion dynamique

	Image thermique $Flag_T P(n) = 0$	**Image thermique** $Flag_T P(n) = 1$
Image de profondeur $Flag_P(n) = 1$	C_G^T et C_T sont mis à 0	Combinaison de tous les coefficients
Image de profondeur $Flag_P(n) = 0$	Ré-échantillonnage des particules	C_G^P et C_D^P sont mis à 0

TABLE 5.4 – Les conditions d'occultations des particules.

les hypothèses suivantes :

$$\text{Si max}\left(C_D^P, C_G^T, C_T\right) < max(C_G^P) \text{ alors } \beta = \frac{1}{2} \text{ et } \alpha = \gamma = \lambda = \frac{1}{6} \tag{5.5}$$

$$\text{Si max}\left(C_D^P, C_G^P, C_T\right) < max(C_G^T) \text{ alors } \gamma = \frac{1}{2} \text{ et } \alpha = \beta = \lambda = \frac{1}{6} \tag{5.6}$$

$$\text{Si max}\left(C_G^P, C_G^T, C_T\right) < max(C_D^P) \text{ alors } \alpha = \frac{1}{2} \text{ et } \beta = \gamma, = \lambda = \frac{1}{6} \tag{5.7}$$

$$\text{Si max}\left(C_G^P, C_G^T, C_D^P\right) < max(C_T) \text{ alors } \lambda = \frac{1}{2} \text{ et } \alpha = \gamma = \gamma = \frac{1}{6} \tag{5.8}$$

Nous avons également remarqué que certaines particules pouvaient sortir en dehors du champ de vision des capteurs. Ce phénomène se reproduit le plus souvent quand la personne est proche des bords des images, comme le montre la Figure 5.11. Cet événement engendre une perturbation de l'étape de mise à jour des poids. Pour cette raison, nous avons ajouté, à chaque instant, deux drapeaux (flag en anglais) pour chaque particule : un au niveau de l'image de profondeur ($Flag_P(n)$ pour la $n$ème particule) et un pour l'image thermique ($Flag_T(n)$). Nous vérifions l'emplacement de chaque particule sur les deux images. Si la position de la particule est dans l'image, nous lui affectons une valeur de 1 au drapeau correspondant et 0 si hors de l'image. Les valeurs de ces drapeaux sont alors utilisées pour adapter la stratégie de mise à jour. Les différentes possibilités avant l'étape de mise à jour sont détaillées dans la Table 5.4. De plus, nous avons choisi de ré-échantillonner toutes les particules si le nombre des particules occultées dépassent la moitié du nombre initial N de particules.

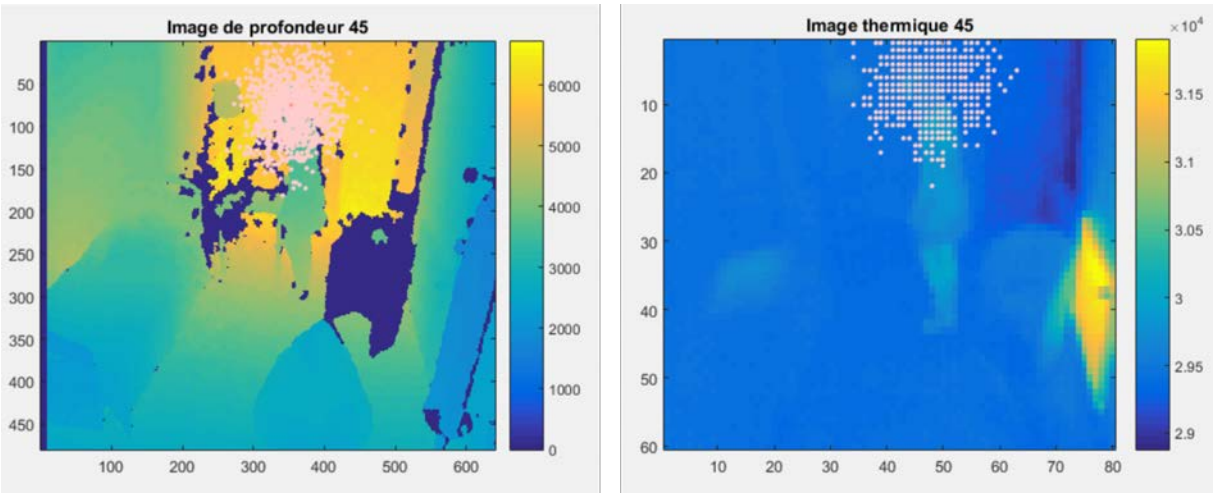


FIGURE 5.11 – Répartition des particules (en blanc) sur : l'image thermique (à gauche) et l'image de profondeur (à droite). Certaines particules présentes dans l'image de profondeur sortent du cadre de l'image thermique.

5. 3.1 Résultats

La Figure 5.12 présente les résultats visuels sur une partie d'une séquence d'images. En regardant l'évolution image par image, nous avons bien constaté que l'importance de chaque coefficient changeait d'une image à une autre. L'utilisation d'un modèle dynamique permet bien d'avoir des résultats plus cohérents le long de la séquence. Pour évaluer la performance de la méthode dynamique, nous avons fait le même test que la méthode statique. La Figure 5.13 montre un exemple de l'amélioration de suivi en appliquant le modèle dynamique par rapport au test C1. Afin de réaliser une évaluation plus quantitative, nous avons également calculé le score de chevauchement et l'erreur de localisation de la position estimée avec la vérité terrain. Ces paramètres sont schématisés par les courbes de succès et de précision (voir Figure 5.14). Ces résultats montrent une nette amélioration apportée par le modèle dynamique par rapport au test le plus performant du modèle statique. Ces résultats sont confirmés par les aires sous la courbe plus élevées pour le modèle dynamique (Table 5.5).

Nous avions également proposé d'améliorer notre méthode dans le cas de mouvements brusques en ajoutant la vitesse instantanée de déplacement de la tête comme paramètre dans le vecteur d'état. Nous constatons que l'ajout du vecteur vitesse permet de mieux gérer les déplacements brusques. La Figure 5.15 montre quelques résultats visuels de cette proposition sur différentes séquences de notre base de données.

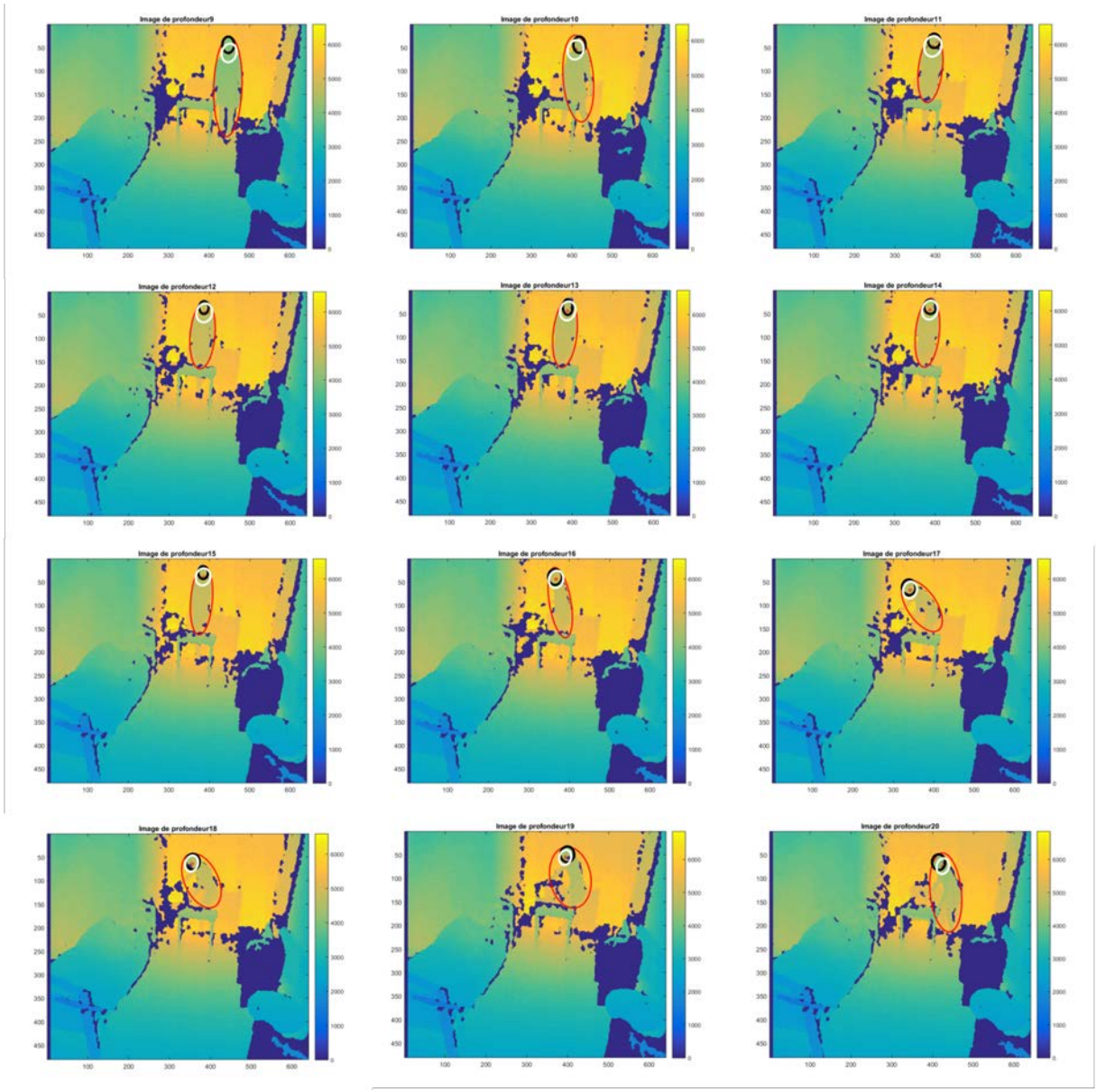


FIGURE 5.12 – Exemple de suivi de la tête de la personne en se basant sur le modèle dynamique. Les résultats du suivi sont en blanc, l'ellipse de la silhouette est en rouge et l'ellipse GT est en noir.

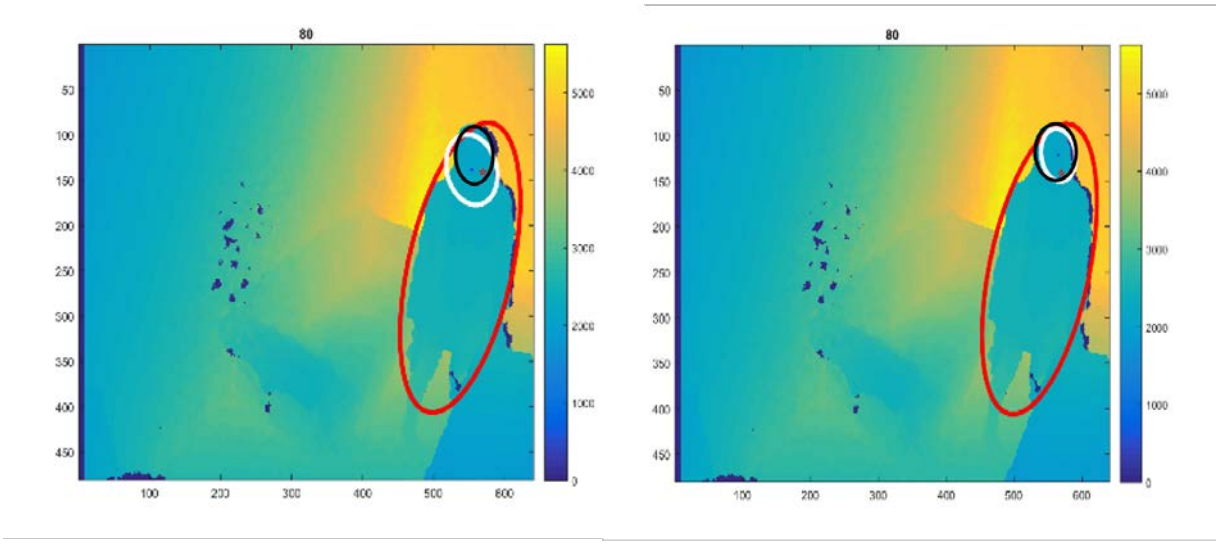


FIGURE 5.13 – Comparaison des résultats du suivi du a) test C1 et b) modèle dynamique. Les résultats du suivi sont en blanc, l'ellipse de la silhouette est en rouge et l'ellipse GT est en noir.

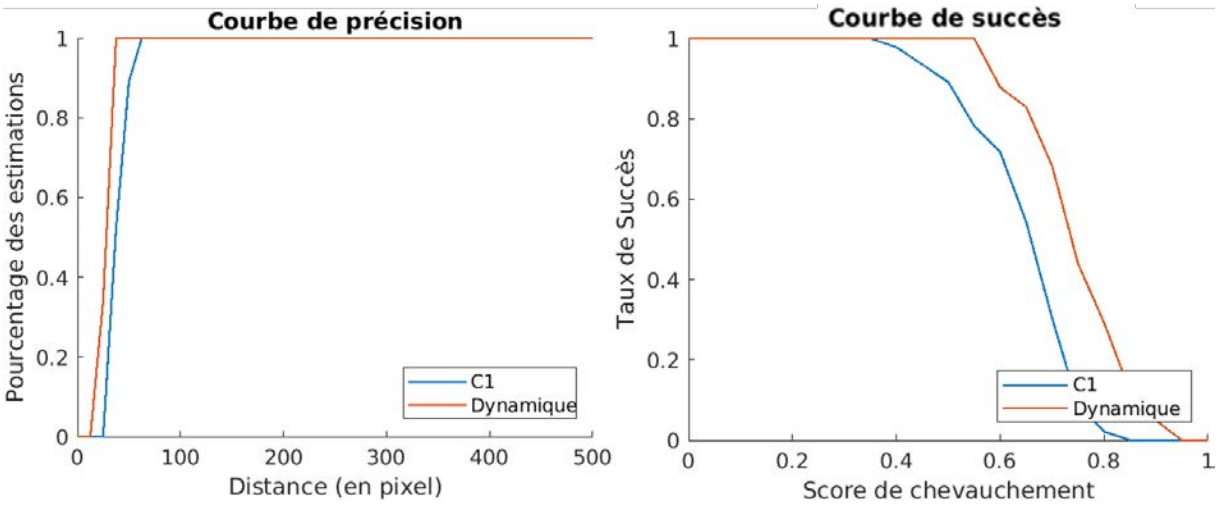


FIGURE 5.14 – Courbes de précision et de succès de l'algorithme de suivi de test C1 et de modèle dynamique.

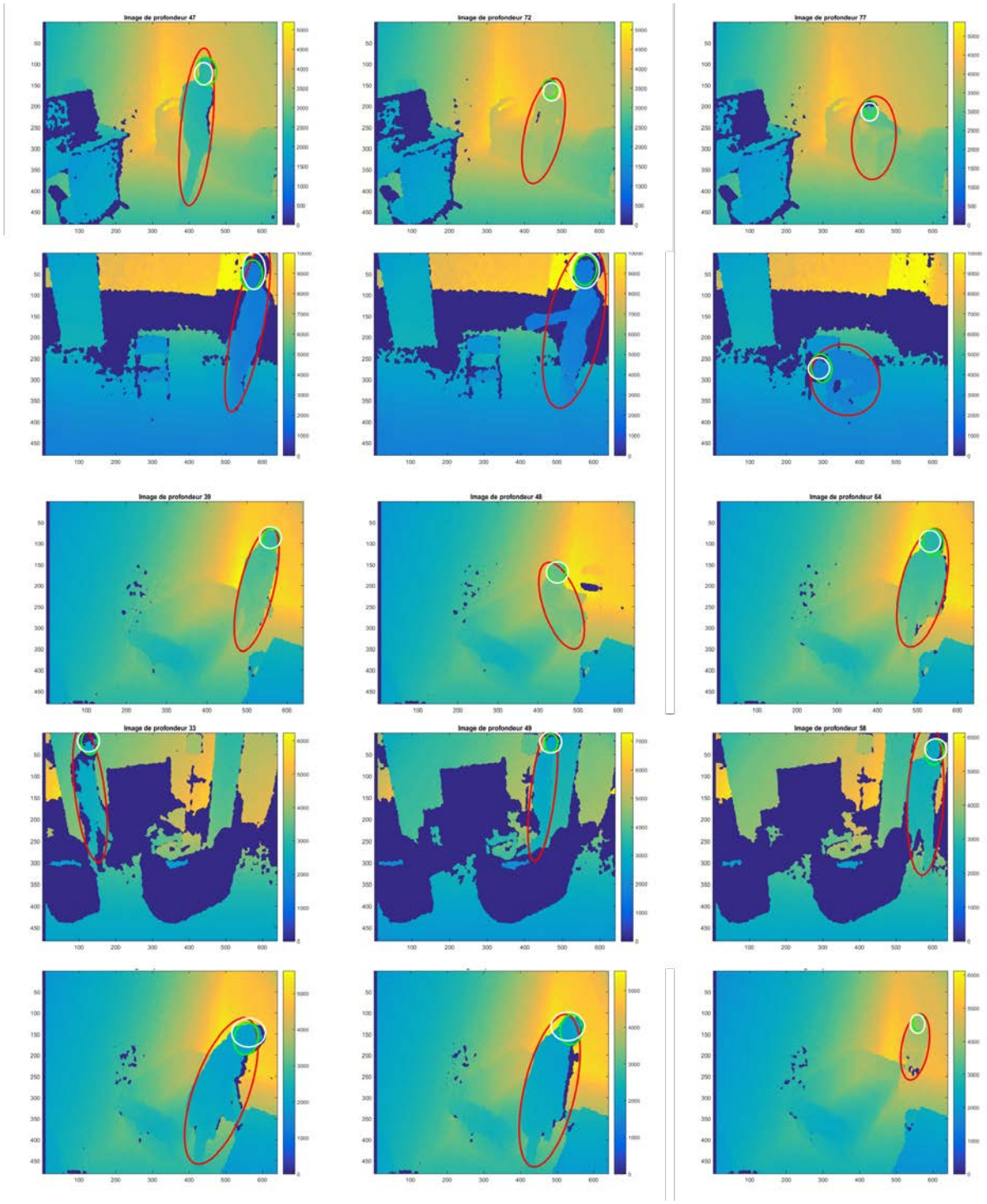


FIGURE 5.15 – Exemples de résultats de suivi d'un modèle dynamique en ajoutant la vitesse dans le vecteur d'état.

	Courbe de précision	**Courbe de succès**
Test C1	0.823	0.619
Modèle dynamique	0.959	0.786

TABLE 5.5 – Aire sous les courbes de précision et de succès moyennes de test C1 et de modèle dynamique.

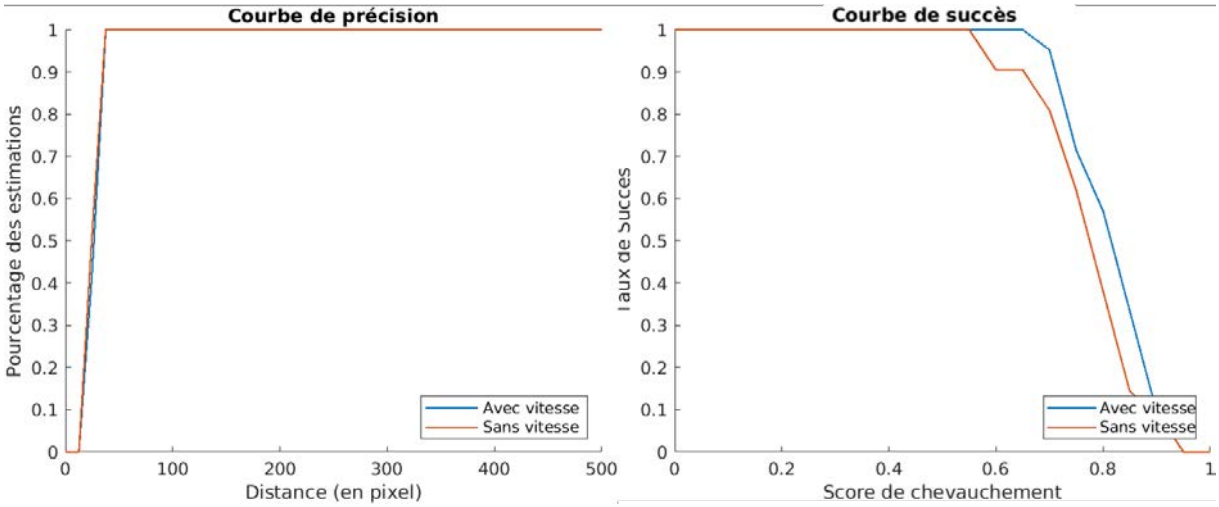


FIGURE 5.16 – Courbes de précision et de succès de l'algorithme de suivi sans et avec l'ajout de la vitesse dans le vecteur d'état.

Nous constatons également cette amélioration sur les résultats quantitatifs donnés par les courbes de précision et de succès (Figure 5.16). Cette amélioration est minime sur la courbe de précision (respectivement AUC de 0.961 et de 0.959 avec et sans vecteur vitesse) mais un peu plus conséquente sur la courbe de succès (respectivement AUC de 0.812 et de 0.786 avec et sans vecteur vitesse). Cette faible différence peut toutefois s'expliquer par le fait que la prise en compte du vecteur vitesse n'améliore le suivi que dans le cas de mouvement brusque. Cette situation est assez peu fréquente dans les séquences et donc l'impact de l'amélioration est assez peu visible dans des métriques globales prenant en compte toutes les images des séquences.

5. 4 Discussion

Dans ce chapitre, nous avons commencé par tester les modèles statiques d'estimation de la tête. Le premier test C1, qui combine les quatre coefficients retenus, fournit

des résultats nettement améliorés par rapport aux modèles de profondeur ainsi que la segmentation seule détaillés dans le chapitre précédent. L'utilisation d'observations thermiques avec celle de profondeur a permis d'améliorer les résultats du suivi. Ensuite, nous avons attribué un facteur d'importance FI à chaque coefficient et nous avons comparé 11 combinaisons différentes de ces coefficients afin d'estimer l'impact de chaque observation. Cependant, les résultats étaient clairement différents entre chaque combinaison en fonction de l'environnement à chaque instant. Pour cette raison, nous avons modifié le FI d'une valeur statique tout au long de la séquence en dynamique (une nouvelle valeur à chaque instant). Cette modification améliore le processus de suivi permettant d'avoir des résultats plus précis. Enfin, nous avons ajouté la vitesse au vecteur d'état pour améliorer l'estimation de la position de la tête, en particulier en cas de mouvement rapide.

5. 5 Conclusion

Nous avons détaillé, dans cette partie, une approche de suivi basée sur un filtrage particulaire en utilisant la fusion des informations thermiques et de profondeur pour détecter la position de la tête d'une personne dans un environnement intérieur. La position, la vitesse, l'orientation et la taille de l'ellipse qui entoure la tête sont utilisées pour prédire la nouvelle position de la tête. Une pondération adaptative a été appliquée sur les coefficients de chaque particule en fonction de l'importance de chaque coefficient pour mettre à jour la position prédite.

Troisième partie

Analyse de la posture de la personne à des fins de prévention de fragilité des personnes âgées

Chapitre 6

Reconnaissance de l'activité de la personne

Introduction

Dans ce chapitre, nous allons dans un premier temps présenter les différents constituants d'un réseau de neurones convolutif. Nous allons ensuite faire une étude bibliographique sur les techniques de reconnaissance de postures des personnes par apprentissage profond. Généralement, ces techniques sont basée sur une première étape de détection/localisation d'objets. Le choix de cette étape est généralement crucial par rapport aux performance du réseau entier en terme de reconnaissance d'activités. Nous allons donc également présenter dans ce chapitre une étude comparative de réseaux dédiés à la détection/localisation d'objets. Cette étude devrait nous permettre de choisir le réseau le plus performant par rapport à notre problématique.

6. 1 L'apprentissage profond (Deep Learning DL)

Ces dernières années, l'apprentissage profond (DL) est utilisé dans la plupart des domaines y compris les technologies de détection des chutes. Islam et al. [124] ont classé les méthodes de détection des chutes basées sur l'apprentissage profond en trois catégories : les systèmes basés sur les réseaux de neurones convolutifs CNN (Convolutional Neural Network en anglais), les systèmes basés sur la mémoire à court et long termes (LSTM) et les systèmes basés sur l'auto-encodeur (la Figure 6.1). De leur côté, Rawat et al. [125] et Voulodimos et al. [126] ont montré que les architectures de CNN sont performantes pour trouver des motifs et des formes dans des images données. En effet, les systèmes de détection de chute qui utilisent plusieurs réseaux basés sur les CNN sont puissants pour détecter et classer des objets. Afin de mieux comprendre les CNN nous allons dans un premier temps détailler les différentes composantes qui les constituent pour la plupart

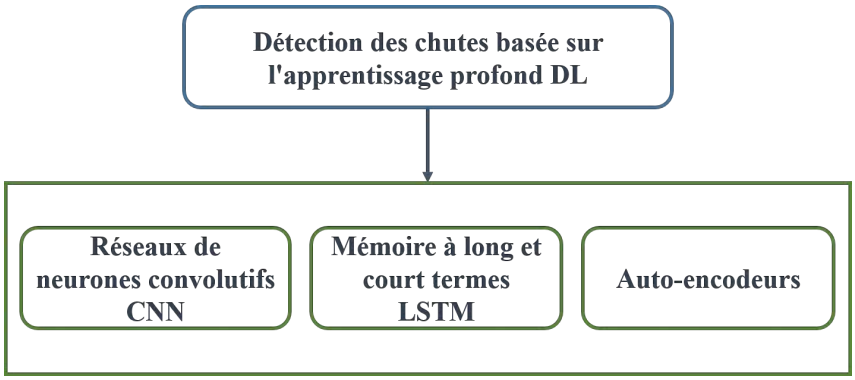


FIGURE 6.1 – Classification des méthodes basées apprentissage profond pour la détection des chutes

d'entre-eux.

6. 1.1 Réseau de neurones convolutif CNN

Un réseau de neurones convolutif (ConvNet/CNN) est un algorithme d'apprentissage profond qui peut prendre en compte une image d'entrée, attribuer de l'importance (poids et biais) à divers aspects/objets de l'image et être capable de les différencier les uns des autres. Le pré-traitement requis dans un ConvNet est beaucoup moins important que pour les autres algorithmes de classification. L'architecture d'un ConvNet est analogue à celle du schéma de connectivité des neurones dans le cerveau humain et s'inspire de l'organisation du cortex visuel. Ce dernier comporte de petites régions de cellules qui sont sensibles à des régions spécifiques du champ visuel. Cette idée a été développée par une expérience de Hubel et Wiesel en 1962 [127], où ils ont montré que certaines cellules neuronales individuelles du cerveau ne réagissent (ou ne se déclenchent) qu'en présence du bord d'une certaine orientation. Par exemple, certains neurones se déclenchent lorsqu'ils sont exposés à des lignes verticales et d'autres lorsqu'ils sont exposés à des lignes horizontales ou diagonales. Hubel et al [127] ont découvert que tous ces neurones sont organisés ensemble pour produire une perception visuelle.

Les neurones individuels ne répondent aux stimulations que dans une région restreinte du champ visuel appelé champ de réception. Un ensemble de ces champs se chevauchent pour couvrir la totalité de la zone visuelle. Un ConvNet est capable de capturer avec succès les dépendances spatiales et temporelles dans une image grâce à l'application de filtres pertinents. L'architecture s'adapte mieux à l'ensemble des données de l'image grâce à la réduction du nombre de paramètres impliqués et à la réutilisation des poids. En d'autres

termes, le réseau peut être entrainé pour mieux comprendre la complexité de l'image. Le rôle du ConvNet est de réduire les images en une forme plus facile à traiter, sans perdre les caractéristiques qui sont essentielles pour obtenir une bonne prédiction.

En général, la convolution d'une image fait référence à une opération d'analyse des caractéristiques d'une image après l'application d'un filtre ou d'un noyau. Une convolution est réalisée en multipliant les valeurs des pixels dans un certain voisinage par une matrice de nombres appelée filtre. Elle fonctionne en déterminant la valeur d'un pixel central en additionnant les valeurs pondérées de tous ses voisins. Le filtre se déplace sur l'image de gauche à droite de manière progressive à la recherche de caractéristiques dans l'image. Les filtres peuvent avoir plusieurs tailles et contenir des motifs différents pour produire des caractéristiques différentes dans l'image [128, 129].

L'architecture de réseau de neurones convolutif est composée de quatre couches principales, comme l'illustre la Figure 6.2 : une couche convolutive, couche de pooling, une couche de flattening et une couche entièrement connectée. Il peut y avoir quelques répétitions de ces couches avant la sortie finale. L'augmentation du nombre de couches rend le réseau plus profond, ce qui peut aider à acquérir d'autres caractéristiques complexes de l'image d'entrée [130].

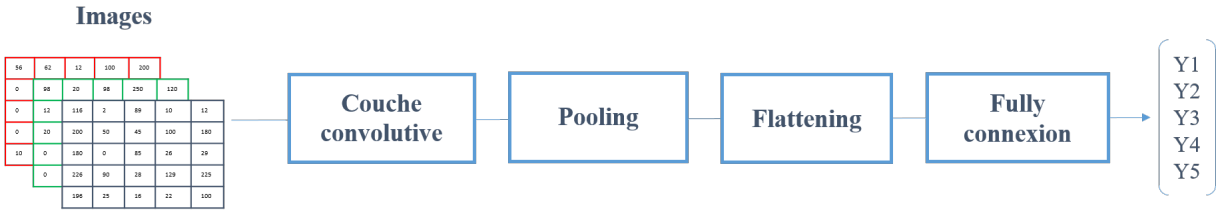


FIGURE 6.2 – L'architecture d'un CNN.

6. 1.1.1 Entrée et sortie du réseau

Á l'entrée du réseau, l'image[1] est traduite par un tableau de valeurs entre 0 et 255 en fonction de la résolution et la taille de l'image. Les valeurs de ce tableau décrivent l'intensité du pixel à ce point. A la sortie, le réseau produit une classe unique ou une probabilité de classes qui décrit mieux l'image, par exemple (0.80 pour un chat, 0.15 pour un chien, 0.05 pour un oiseau, etc).

1. Actuellement les réseaux prennent des images sur 8 bits

6. 1.1.2 Couche convolutive

Le réseau peut procéder à la classification des images en recherchant des caractéristiques de bas niveau telles que les bords et les courbes, puis en élaborant des concepts plus abstraits par le biais d'une série de couches convolutives. Comme le montre la Figure 6.3, chaque couche convolutive subit trois opérations principales : convolution, normalisation spatiale par lots et unité linéaire rectifiée (Rectified Linear Unit en anglais ReLU).

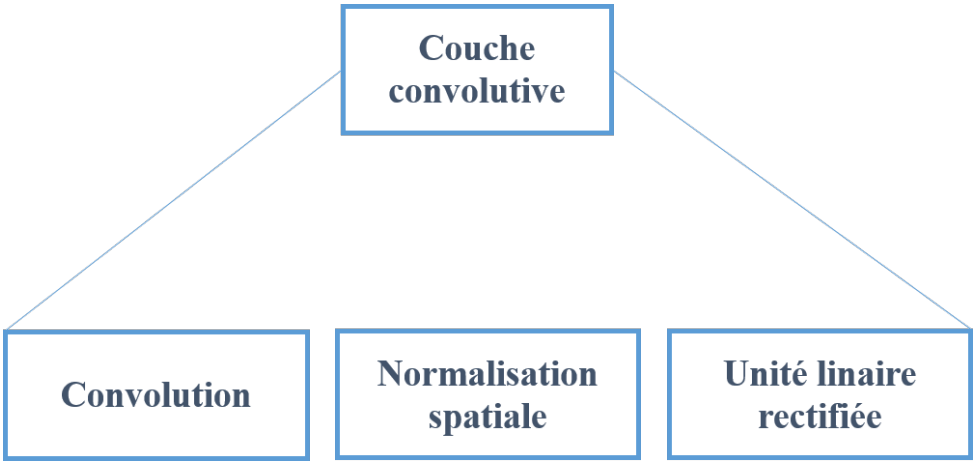


FIGURE 6.3 – Couche convolutive

6. 1.1.2.1 Convolution. Cette étape consiste à extraire des caractéristiques de l'images d'entrée. La convolution préserve les relations spatiales entre les pixels en apprenant les caractéristiques de l'image à l'aide d'un filtre/noyau (Figure 6.4). Comme nous l'avons expliqué, chaque image peut être considérée comme un tableau de plusieurs dimensions.

Considérons une image 5×5 dont les valeurs des pixels sont entre 0 et 255 (Figure 6.5). La matrice formée en faisant glisser le filtre sur l'image d'entrée et en calculant le produit des points est appelée "caractéristique convoluée" ou "carte d'activation" ou "carte des caractéristiques". La première convolution est illustrée sur la Figure 6.5. Il est important de noter que les filtres agissent comme des détecteurs de caractéristiques à partir de l'image d'entrée originale. Le résultat du parcours de ce filtre sur toutes les valeurs de la matrice est affiché par la Figure 6.6 [131]. Cette étape dépend des paramètres de profondeur de la couche (Depth), de pas (Stride) et de la marge de zéro (Zero padding) :

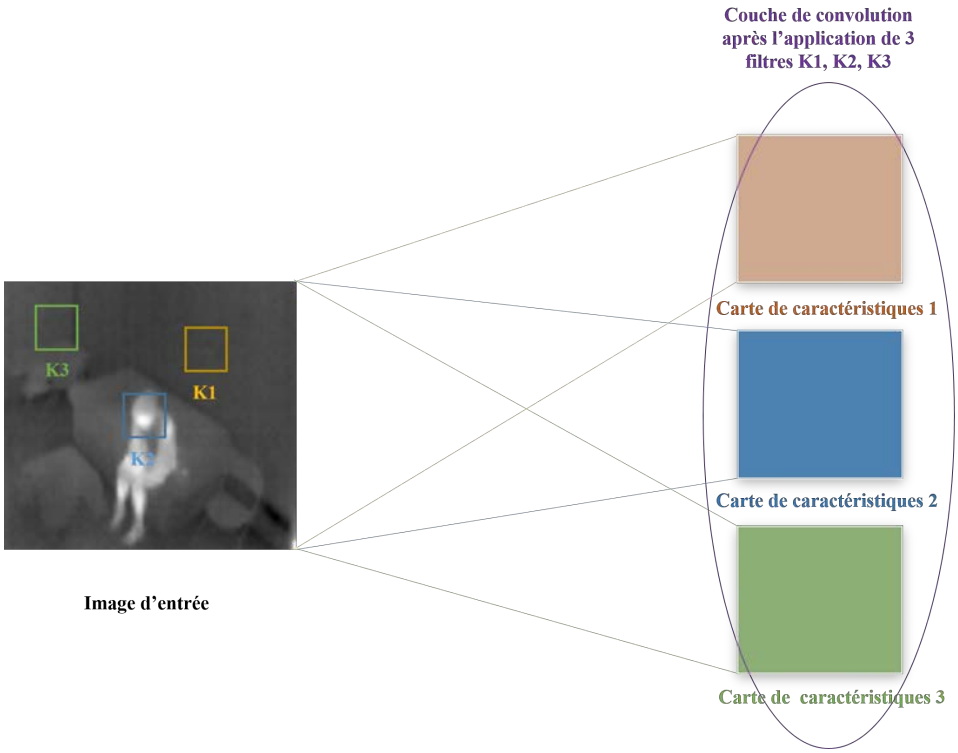


FIGURE 6.4 – Première couche convolutive après l'application de trois filtres sur l'image d'entrée.

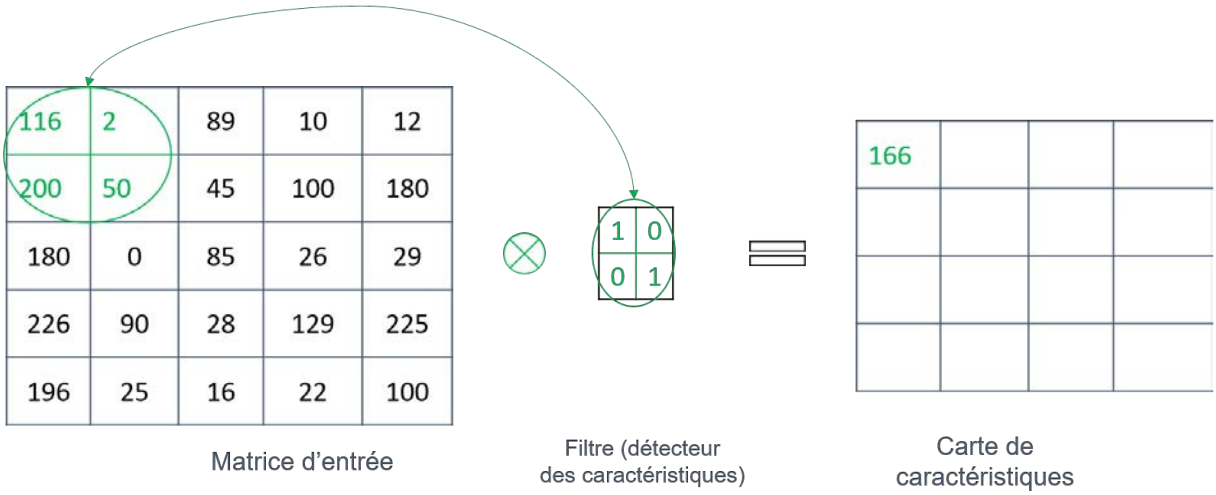


FIGURE 6.5 – Étape de convolution.

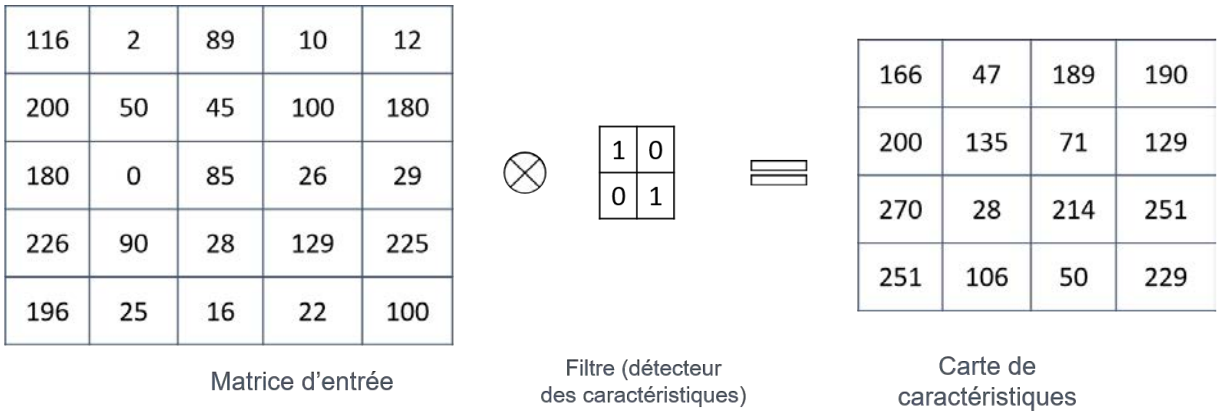


FIGURE 6.6 – Résultats de convolutions avec un filtre de dimension 2.

6. 1.1.2.1.1 Profondeur de la couche (Depth). Il s'agit du nombre de filtres utilisés sur un même champ récepteur. Plus le nombre est élevé plus les résultats sont proches à la réalité. Par exemple, la Figure 6.7 montre que quatre filtres ont été appliqués sur l'image, ce qui permet de produire quatre cartes de caractéristiques. La région encadrée en rouge est représentée par une région sur chacune de ces cartes de caractéristiques. Au total, elle est donc représentée par quatre régions.

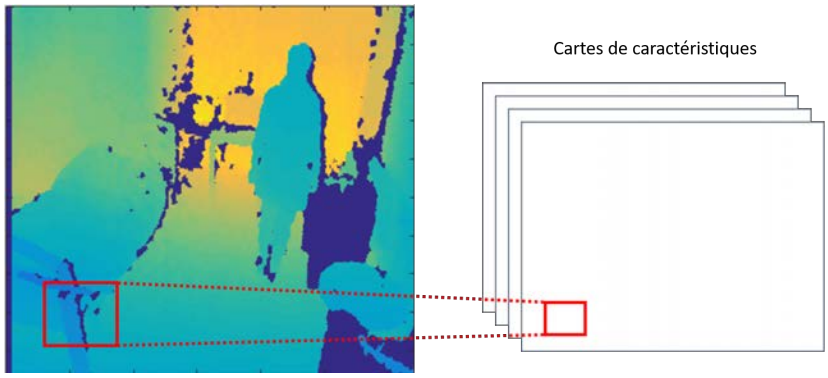


FIGURE 6.7 – Convolution de profondeur 4.

6. 1.1.2.1.2 Pas (Stride). Le stride est le pas de déplacement du filtre sur la matrice d'entrée. Lorsque le Stride est égale à 1, tous les pixels sont parcourus par les filtres. Quand le Stride est fixé à 2, les filtres sautent d'un pixel lors du balayage de l'image. Plus le Stride est grand, plus les cartes obtenues sont petites.

6. 1.1.2.1.3 Marge de zéro (Zero padding). Il est pratique d'ajouter des zéros autour de la matrice d'entrée afin d'appliquer les filtres sur tous les pixels de l'image y compris ceux du bord et de respecter le format de l'image à l'entrée (Figure 6.8). Le principal avantage de cette opération est de contrôler la taille des cartes de caractéristiques. L'ajout de zéros est également appelé convolution large, sinon c'est une convolution étroite [132].

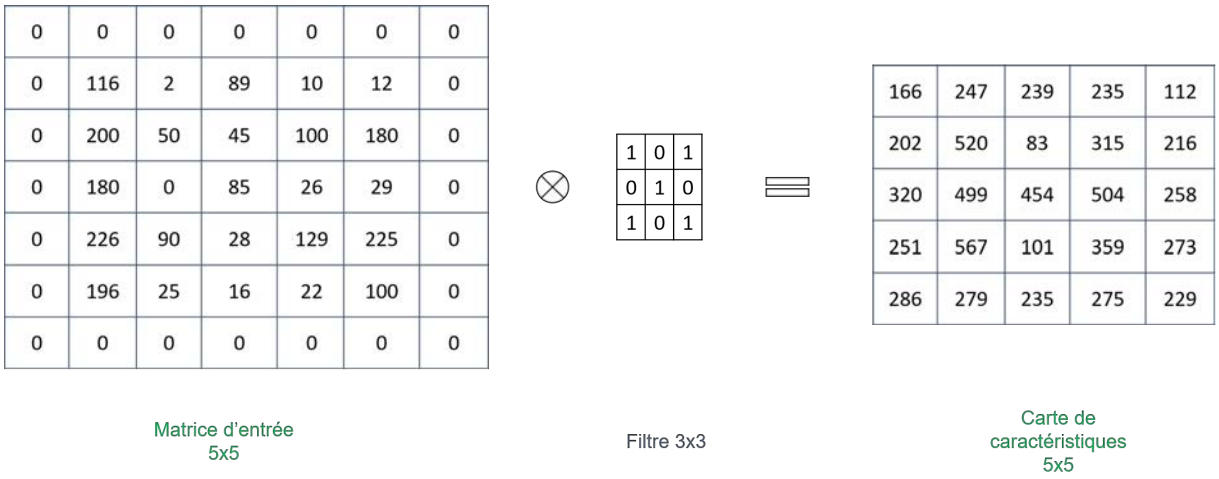


FIGURE 6.8 – Étape d'ajouter des zéros sur le bord de l'image.

6. 1.1.2.2 Normalisation spatiale par lot. Une étape de normalisation est appliquée après chaque convolution puisque les images peuvent avoir des formats différents. Comme l'exemple de notre base de données, nous avons des valeurs de pixels thermiques très faibles et d'autres de profondeur très élevées. La normalisation spatiale par lots calcule la moyenne et l'écart-type à partir d'un mini lot d'entrée. Ensuite, toutes les valeurs de la carte de caractéristiques sont soustraites par la moyenne calculée et divisées par l'écart type calculé :

$$y = \frac{x - Mean(x)}{\sigma + \beta} \tag{6.1}$$

avec σ est l'écart type et β est une constante égale à 10^{-3} pour éviter le cas où $\sigma = 0$. La soustraction de chaque valeur avec la moyenne est faite pour centrer les données autour de zéro. Ensuite, elles sont normalisées en les divisant par l'écart-type. L'objectif de cette normalisation est de mettre en échelle les valeurs d'entrée pour qu'elles puissent avoir la

même importance en phase d'apprentissage [132].

6. 1.1.2.3 Unité linéaire rectifiée (ReLU). Afin d'améliorer l'efficacité du traitement, une opération supplémentaire appelée fonctions d'activations est utilisée après chaque normalisation. Ces fonctions sont utilisées pour transformer le domaine d'une entrée en un autre différent, par exemple pour passer de la linéarité à la non-linéarité. Ceci permet au réseau d'avoir des propriétés distinctes de l'entrée pour l'apprentissage [132]. Sa sortie est donnée par l'équation 6.2 :

$$X_{out} = Max(0, X_{in}) \tag{6.2}$$

ReLU est appliquée par pixel et elle remplace par zéro toutes les valeurs négatives de la carte de caractéristiques. Le but de ReLU est d'introduire la non-linéarité dans le ConvNet.
De même, certaines autres fonctions d'activation sont utilisées dans les réseaux de neurones : la fonction sigmoïde (sigmode) $(f(x) = (1 + /exp(-x))^{-1})$ qui associe une entrée à une valeur comprise entre 0 et 1 et la fonction tangente hyperbolique (Tanh) $(f(x) = \tanh(x))$ qui associe une entrée à une valeur comprise entre -1 et 1 [132].

6. 1.1.3 Couche de Pooling

C'est la dernière étape de sélection des caractéristiques. Elle réduit la dimension de chaque carte de caractéristiques en conservant les informations les plus importantes. L'opération de Pooling peut être réalisée sous plusieurs formes : Max, Moyenne, Somme, etc.
Dans le cas d'un Max Pooling, nous définissons un voisinage spatial (par exemple, une fenêtre 2×2) et nous prenons le plus grand élément de la carte de caractéristiques rectifiées dans cette fenêtre. Au lieu de prendre l'élément le plus grand, nous pouvons également prendre la moyenne (Average Pooling) ou la somme de tous les éléments de cette fenêtre [132]. La Figure 6.9 montre un exemple de Max Pooling sur une carte des caractéristiques rectifiées (obtenue après une opération de convolution et de ReLU) en utilisant une fenêtre 2×2.

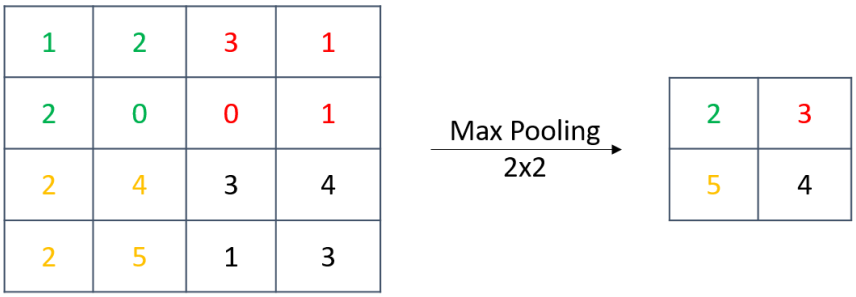


FIGURE 6.9 – Exemple de Max Pooling.

6. 1.1.4 Flattening

C'est l'opération qui consiste à convertir tous les tableaux bidimensionnels résultants en un seul long vecteur linéaire continu, comme le montre la Figure 6.10

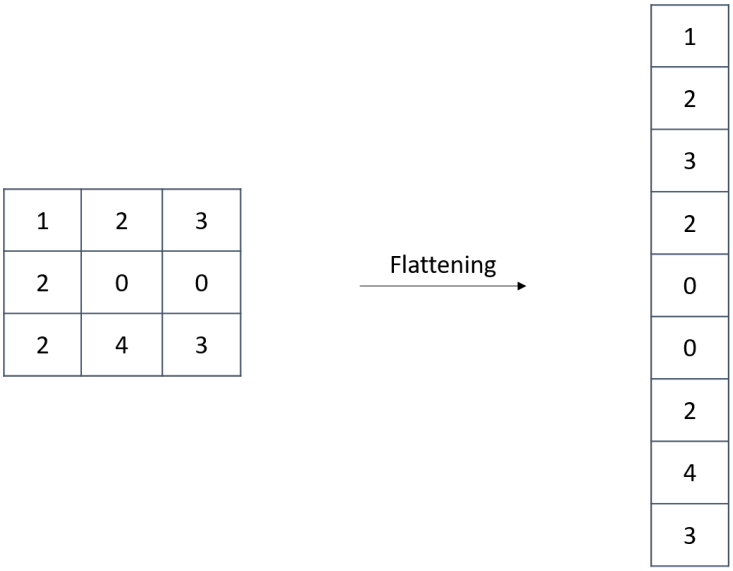


FIGURE 6.10 – Exemple de Flattening.

La sortie du Flattening servira alors d'entrée à un réseau de neurones artificiel, le fully connected layer le plus souvent, qui fera le raisonnement haut niveau.

6. 1.1.5 Fully connected layer FC

La couche entièrement connectée (Fully connected layer FC en anglais) est un perceptron multicouche qui utilise une fonction d'activation softmax en sortie. Les sorties

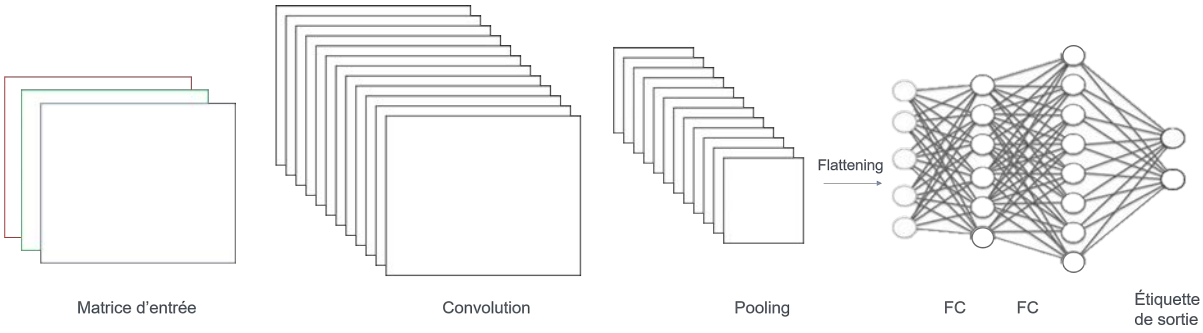


FIGURE 6.11 – Processus d'un CNN.

des couches convolutives et de la couche Pooling représentent des caractéristiques de haut niveau de l'image d'entrée. L'objectif de la couche Fully connected est d'utiliser ces caractéristiques pour classer l'image d'entrée dans différentes classes en fonction de l'ensemble de données d'apprentissage. Les caractéristiques, après être passées par plusieurs convolutions et Pooling, sont empilées ensemble dans la couche FC. Toutes ces caractéristiques sont connectées à la sortie. La couche FC est connectée d'une manière linéaire. Cependant, elle permet une combinaison non linéaire de ces caractéristiques pour extraire des caractéristiques plus complexes en utilisant plusieurs couches entièrement connectées [132].

6. 1.1.6 Sortie du réseau

La dernière couche est la couche de sortie qui prédit la classe d'appartenance. Cette couche de sortie utilise un classifieur appelé softmax qui calcule la probabilité d'appartenance à chaque classe à partir des scores c de la couche FC. Le softmax σ calcule l'exponentiel de ces scores et divise ensuite chacune de ces valeurs par la somme pour normaliser et obtenir une distribution uniforme des probabilités qui s'additionnent à 1, comme la montre l'équation 6.3.

$$\sigma_{c_i} = \frac{e^{c_i}}{\sum_i c_i} \tag{6.3}$$

Un log négatif est alors appliqué à ces probabilités normalisées pour calculer la perte d'entropie croisée. Par conséquent, le logarithme de softmax qui est également appelé LogSoftMax, pousse la probabilité logarithmique normalisée de la classe correcte pour atteindre la valeur de 1 [133]. Au final la valeur maximale parmi ces probabilités représente

l'étiquette de classe de sortie.

6. 1.2 Reconnaissance de postures des personnes par apprentissage profond

La classification des postures est le principal objectif de ma thèse puisque cette classification peut fournir un indice sur la fragilité de la personne âgée ainsi que la détection des chutes. Des travaux récents ont traité cette thématique en utilisant différents capteurs d'acquisition des données (accéléromètres, radar et caméra). Dans cette partie, nous nous intéressons uniquement aux méthodes de reconnaissance de posture par apprentissage appliquées sur des images.

Adhikari et al [134] ont proposé un système de détection des chutes basé sur des images RGB-D. Ils ont appliqué des CNN pour reconnaître les activités de vie de la personne âgée ainsi que les chutes. Ils utilisent leur propre base de données provenant de différents environnements intérieurs. Ils ont enregistré plusieurs activités par différentes personnes. L'ensemble des données contient un total de 21499 images. Une répartition de 73% et 27% a été effectuée respectivement pour la création de la base d'apprentissage et de test. La précision globale du système proposé est de 74%. La sensibilité est de 99% lorsque la personne est en position couchée. Cependant, le système atteint une sensibilité très faible lorsque la personne est assise ou penchée. Le système a été développé pour un scénario d'une seule personne. Lima et al. [135], Chen et al. [136], Li et al. [137] ainsi que Simonyan et al. [138] ont proposé un système de reconnaissance de l'activité humaine basé sur le CNN qui extrait des caractéristiques spatiales et temporelles. Jain et al. [139], Rafi et al. [140] et Luo et al. [141] ont également proposé une architecture de réseau convolutif multicouche. Ce réseau apprend les caractéristiques de bas niveau et un modèle spatial de haut niveau pour estimer la posture humaine. Jain et al. ont estimé que la classification des postures peut être un indice de la fragilité de la personne âgée [139].

De même, Nunez-Marcos et al [142] ont proposé une détection des chutes basée sur des CNN où ils classent les chutes et les autres activités de la vie quotidienne à partir de la technique d'apprentissage par transfert. Ils ont utilisé le modèle CNN VGG-16 entrainé sur Imagenet. Les flux optiques représentent le mouvement entre deux images consécutives. Cependant, cette information de mouvement est trop courte pour représenter une chute. Nunez-Marcoz et al. ont alors utilisé un lot de 20 images comme entrée du réseau CNN. Li et al [143] ont présenté une méthode de classification des activités quotidiennes de

la personne ainsi que de la chute. La base de données "UR Fall Detection" (cf. section 2.4.1.1) a été utilisée dans ce système. Le CNN est appliqué sur chaque image. Une base de 10 dossiers contenant 850 images chacun de 'UR Fall Detection" a été utilisée pour évaluer la performance du système. Les performances de cette approche ont obtenu une sensibilité de 100% et une précision de 99,98%. Comme la base de test utilisée a le même arrière-plan, les mêmes couleurs et le même environnement, les changements d'arrière-plan peuvent dégrader la performance du système. Cependant, la performance n'a pas été mesurée pour d'autres environnements.

Une autre méthode de détection de chute a été proposée par Yu et al [144]. Ces chercheurs ont segmenté les images d'entrée en utilisant la technique de soustraction du fond pour extraire la silhouette du corps humain. Dans le système proposé, le CNN est appliqué sur les silhouettes extraites qui correspondent à des mouvements humains comme se tenir debout, s'asseoir, se pencher et s'allonger. Un ensemble de données de posture personnalisé contenant 3216 postures (804 debout, 769 assis, 833 courbés et 810 couchés) a été utilisé pour tester le système. Cette proposition atteint une précision égale à 96,88%. Shojaei-Hashemi et al [145] ont exploré une autre approche d'apprentissage profond en utilisant les caractéristiques du squelette 3D obtenues à partir du SDK Microsoft Kinect pour alimenter un réseau long short-term memory (LSTM) qui permet d'intégrer des données temporelles afin d'identifier les chutes. Ils ont mentionné que les images de chutes sont limitées par rapport à celles d'activités normales comme la marche. Par conséquent, ils entraînent d'abord un LSTM multiclasse sur des activités quotidiennes de la personne âgée et ils transfèrent le poids appris pour la dernière couche d'un autre LSTM à deux classes. Ils ont utilisé la technique d'apprentissage par transfert pour éviter d'avoir besoin d'une énorme base de données de chute. Ils ont réussi à obtenir une valeur de précision de 0,9323 et une valeur de rappel de 0,9612 sur l'ensemble des données de reconnaissance d'actions RGB+D.

Solbach et al. [146] ont proposé une approche de détection des chutes basée sur la vision pour les personnes âgées, qui combine des informations de profondeur avec une estimation de la pose humaine en 2D basée sur un CNN pour estimer les points caractéristiques d'une personne en 3D. Ces points sont ensuite utilisés pour obtenir le CoG (centre de gravité) des points détectés et le point le plus haut du corps. En utilisant la distance euclidienne, la distance entre deux points dérivés et les points 3D du plan de sol sont alors mesurés. Ils considèrent que la personne est tombée uniquement lorsque la distance avec le CoG est inférieure à 0,7 m. Leur approche a obtenu un taux de vrai positif égale à 0,933 dans

l'environnement intérieur. Cependant, l'approche se base sur la détection du sol, ce qui peut affecter la précision lorsqu'aucune information de profondeur n'est disponible, ce qui peut être simplement dû aux réflexions. Ils ont créé leur propre ensemble de données pour analyser la chute. Quanzeng et al [147] ont proposé une approche de détection des chutes en deux étapes. La première partie est basée sur PCANet, un réseau pour prédire l'étiquette de chaque image après soustraction de fond. Ils ont ensuite utilisé le SVM, dans la deuxième partie, pour classer les prédictions issues de PCANet en deux classes : chute ou pas chute.

6. 2 Méthodes basées sur les réseaux convolutifs CNN

Les méthodes de détection d'objets basées sur les réseaux convolutifs *CNN* peuvent être classés en deux catégories (Figure 6.12).

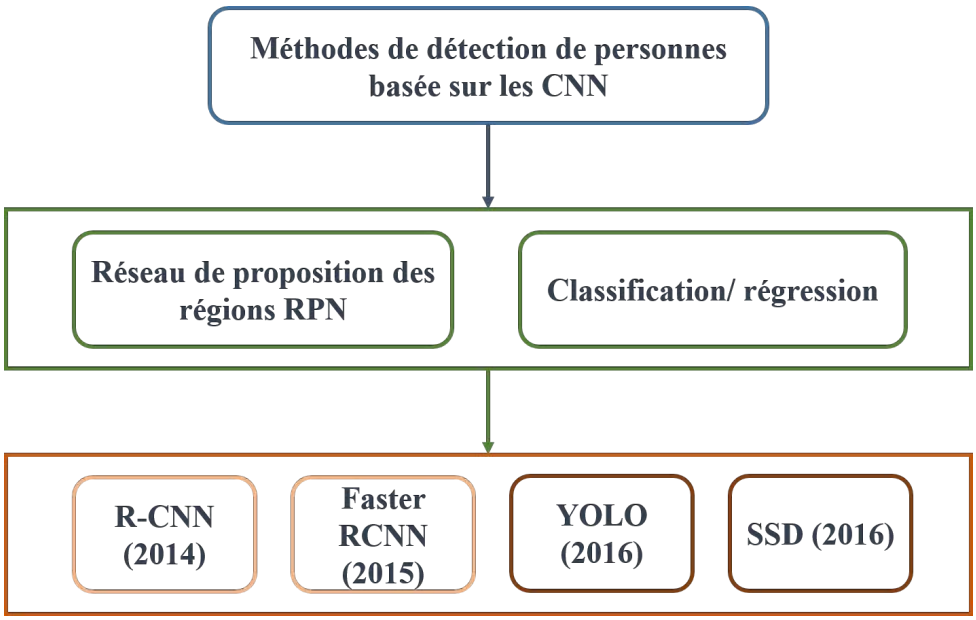


FIGURE 6.12 – Méthodes de détection basées sur les réseaux CNN.

La première catégorie permet de localiser les objet à l'aide d'un réseau de propositions des régions (*RPN*) en premier temps puis de classer chaque proposition dans différentes catégories d'objets. L'autre stratégie de méthodes, plus récente, considère la détection d'objets comme un problème de régression/classification, en adoptant une boîte englobante unitaire pour obtenir directement les résultats finaux (catégories et emplacements).

Les méthodes basées sur les réseaux de propositions de régions sont *R-CNN* [148], *SPP-net* [149], *Fast R-CNN* [150], *Faster R-CNN* [151],*R-FCN* [152], *FPN* [153] et *Mask R-CNN* [154]. Les méthodes basées sur la régression/classification sont *MultiBox* [155], *G-CNN* [156], *YOLO* [157], *SSD* [158], *YOLOv2* [159]. Nous avons choisis de comparer les méthodes les plus connues : *R-CNN*, *Faster R-CNN*, *YOLO*, *YOLOv2*, *SSD300* et *SSD512*, détaillées ci-dessous.

6. 2.1 Méthodes basées sur les réseaux RPN

Les méthodes basées sur les réseaux *RPN* sont composées de deux réseaux montés en étage. Elles permettent d'obtenir un aperçu sur l'ensemble de la scène dans un premier temps, puis se concentrer sur les régions d'intérêt. Girshick et al. ont proposé, en 2014, le modèle *R-CNN* [148], qui utilise le réseau *CNN* pour prédire les zones d'intérêt (Figure 6.13). Dans un premier temps, un générateur de propositions de région est exploité pour

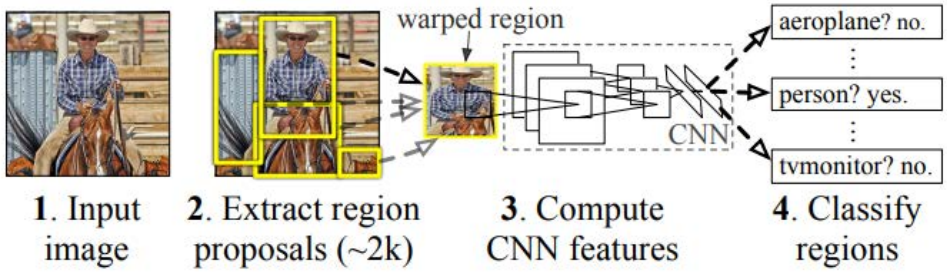


FIGURE 6.13 – Architecture de la méthodes de détection d'objets R-CNN [148].

détecter les régions d'intérêt. Ce générateur utilise la technique de la recherche sélective [160] pour générer environ 2000 propositions de régions pour chaque image. Au niveau de l'étape d'extraction de caractéristique, chaque proposition est recardé et un réseau *CNN* est appliqué pour extraire une représentation finale pour chaque région. Finalement, les régions sont classées par un *SVM* linéaire prédéfini pour plusieurs classes. Malgré les avantages de cette méthode par rapport aux méthodes traditionnelles, l'apprentissage reste gourmand en temps et en mémoire. Plusieurs améliorations ont été effectuées sur ces modèles de détection d'objets, Ren et al. [151] ont introduit un réseau de proposition de région (*RPN*) supplémentaire, qui agit d'une manière rapide par rapport au générateur de régions.

6. 2.2 Méthodes basées sur la régression/classification

Les méthodes basées sur les réseaux de proposition de régions sont composés de plusieurs étapes corrélées : la génération de propositions de régions, l'extraction de caractéristiques à l'aide d'un réseau *CNN*, la classification et la régression par boîte englobante. Ces étapes sont généralement formées séparément. Même dans le récent "end to end" module de cette catégorie (*Faster R-CNN*), une formation alternative est encore nécessaire pour obtenir des paramètres de convolution partagés entre le *RPN* et le réseau de détection. Par conséquent, le temps de manipulation de différentes étapes reste un inconvénient pour les applications en temps réel. Concernant les systèmes de régression/classification, le principe est différent. Ces méthodes configurent chaque pixel en l'affectant un emplacement dans la boîte englobante et une probabilité de classe, ce qui réduit le temps de calcul.

Redmon et al. ont proposé le *YOLO* [157] en 2016, qui utilise les dernières cartes de caractéristiques pour prédire la localisation et la classification de chaque boîte englobante. Le principe de cette méthode est illustré dans la figure 6.14. L'image est divisée en une grille ($S \times S$)et chaque cellule contient une information sur l'objet d'intérêt centré dans cette cellule. Autrement dit, chaque cellule contient plusieurs boîtes englobantes de différentes formes et chaque boîte a un score d'appartenance de cet objet aux classes prédéfinies en entrée.

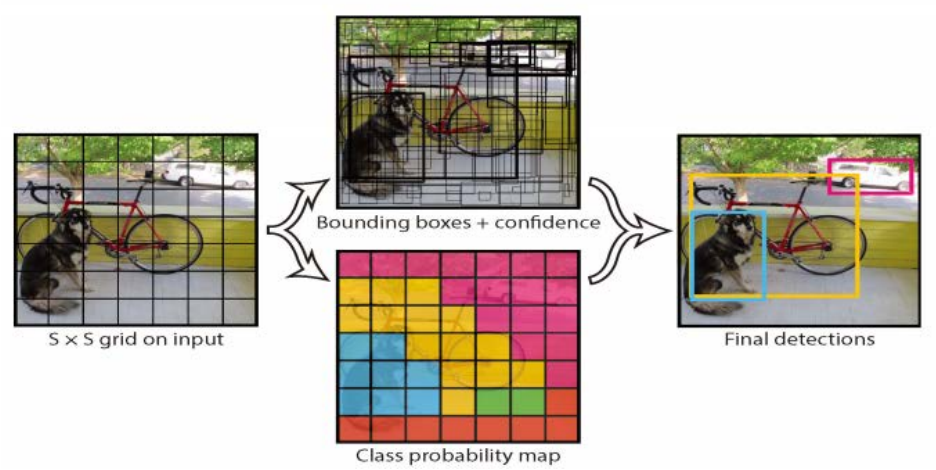


FIGURE 6.14 – Principe de la méthode YOLO [157].

Ensuite, d'autres versions ont été proposées qui améliorent la version basique. Prenons

l'exemple de *YOLOv2*, il limite le nombre de boîtes englobantes par cellule et modifie l'entrainement en multi-échelles. Par contre, *YOLO* a des difficultés à gérer les objets de petites tailles, ce qui est dû aux fortes contraintes spatiales imposées aux prédictions de la boîte englobante [157]. Pour résoudre ce problème, Liu et al. ont proposé la méthode *SSD* (Single Shot MultiBox Detector) [158] (détaillée dans le chapitre suivante). Cette méthode est plus avantageuse que le *YOLO* puisque elle est basée sur un ensemble de boîtes englobantes par défaut avec différents rapports de longueur et largeur définis à l'avance. Pour traiter des objets de tailles diverses, le réseau fusionne les prédictions de plusieurs cartes de caractéristiques avec différentes résolutions. Il existe deux variantes de SSD (*SSD300* et *SSD512*) déterminées par la taille d'image d'entrée (300×300 ou 512×512 respectivement).

La Table 6.1 illustre une comparaison de ces méthodes sur la base de données *VOC07*. Nous constatons que la performance la plus élevée de toutes les classes est celle de *YOLOv2* et le temps de calcul le plus rapide est celui de *SSD300* et *YOLO*. Puisque nous nous inté-

Méthode	mAP(%)	Temps de test (sec/img)	Fréquence (FPS)
R-CNN	66,0	32,84	0,03
Faster R-CNN	73,2	0,11	9,1
YOLO	63,4	**0,02**	45
YOLOv2	**78,6**	0,03	40
SSD300	74,3	**0,02**	**46**
SSD512	76,8	0,05	19

TABLE 6.1 – Comparaison de différentes méthodes de détection d'objets sur la base de données *VOC07*.

ressons à la classe "personne" uniquement, nous avons présenté les performances de cette classe sur les bases de données *VOC07* et *VOC12*. Nous observons que les performances sont plus élevées pour la méthode *SSD* (Table 6.2).

La performance et le temps de calcul sont les critères les plus importants pour notre projet. Ainsi, Nous avons choisi de baser notre système de reconnaissance de postures sur la méthode SSD. De plus, la résolution des images thermiques est faible (80×60) d'où le choix de la première variante. Dans le chapitre suivant, nous présenterons la méthode *SSD300* plus en détails ainsi que notre système de reconnaissance de postures.

Méthode	VOC 2007	VOC 2012
R-CNN	58,7	57,8
Faster R-CNN	76,7	79,6
YOLO	N.A	63,5
YOLOv2	N.A	81,0
SSD300	81,4	85,6
SSD512	**84,6**	**88,6**

TABLE 6.2 – Comparaison de différentes méthodes de détection d'objets sur les bases de données *VOC07* et *VOC12* de la classe "personne" uniquement.

6. 3 Conclusion

Dans ce chapitre, nous avons étudié des méthodes de classification d'activités en s'appuyant sur les réseaux de neurones convolutifs (CNN). Nous avons détaillé l'architecture de ces réseaux que nous avons utilisée pour développer notre système de reconnaissance des posture, détaillé dans le chapitre suivant.

Chapitre 7

Classification de postures des personnes par des méthodes de Deep Learning

Introduction

La classification des postures est le deuxième objectif de nos travaux de recherche car elle fournit des indices sur la fragilité de la personne âgée ainsi que la détection des chutes. En effet, l'emplacement du corps et le type de postures permettent de différencier les activités normales et les chutes. Dans ce chapitre nous décrivons la méthode de reconnaissance de postures appliquée sur les images de profondeur et thermique. Tout d'abord, nous représentons le type de données à traiter et les contraintes qu'elles imposent. Ensuite, nous détaillons la technique que nous avons utilisée pour fusionner les images de profondeur et thermiques. Enfin, nous définissons la méthode de reconnaissance des postures ainsi que les expérimentations faites.

La Figure 7.1 définit les différentes étapes de la méthode de classification de postures que nous avons développée. La première étape consiste à acquérir une base de données qui contient 1300 exemples de postures à détecter. Le choix des données d'apprentissage est très important pour avoir un système de classification précis et robuste. Il est conseillé d'avoir plusieurs postures de différentes personnes dans plusieurs pièces de styles différents. Puis nous avons annoté ces données manuellement et avons encadré toutes les postures par une boîte englobante. Ensuite, ces données sont traitées, en augmentant artificiellement (transformations géométriques ou d'éclairage par exemple) le nombre des images, et apprises à partir d'une méthode d'apprentissage profond. La base de données est ensuite divisée en trois parties : une base d'apprentissage, une base de validation et une base de test. Enfin, nous testons les résultats de classification sur la base de test et évaluons les performances de cette méthode.

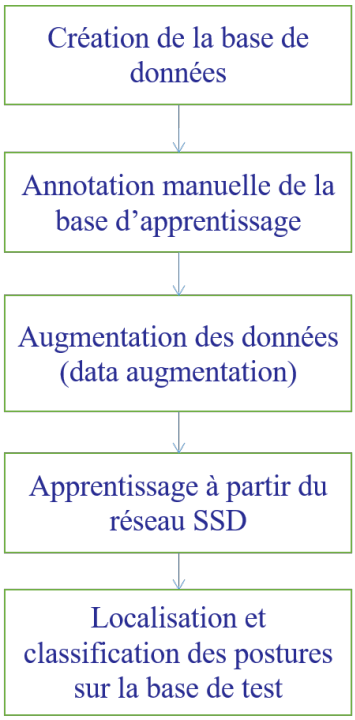


FIGURE 7.1 – Procédure de classification des postures

7. 1 Base de données

Nous avons détaillé dans la section 2.4.2 notre méthode de création de la base de données. La Figure 7.2 montre l'état brut des images après acquisition. Nous rappelons que les données enregistrées sont des images 16 bits de tailles (640×480) et (80×60) pour le capteur de profondeur et thermique respectivement. Nous rappelons que ces images ont des champs de vision légèrement différents. Lors de la campagne d'acquisition, nous avons choisi de fixer les capteurs de manière à avoir le même champ de vision. Cependant, les angles d'ouverture du capteur thermique sont plus petits que ceux du capteur de profondeur. Donc les images thermiques ont moins d'information sur la pièce par rapport aux images de profondeur.

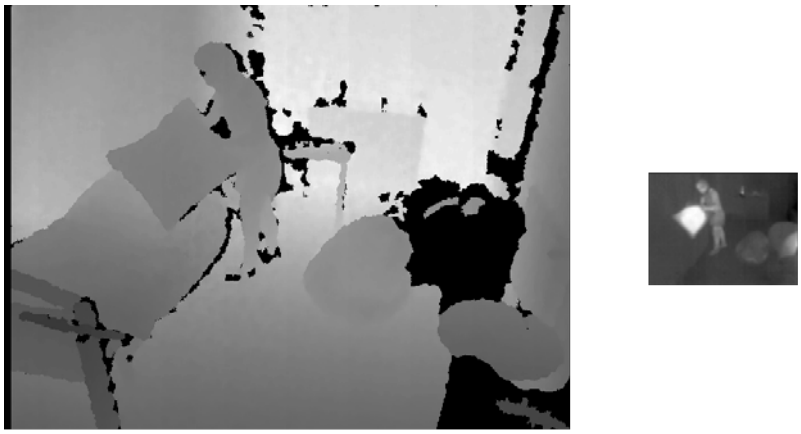


FIGURE 7.2 – Images de profondeur (a) et thermique (b).

7. 1.1 Compensation de la perte d'information des images de profondeur

Nous avons constaté qu'au niveau d'une image de profondeur certains pixels peuvent avoir une valeur nulle (présence des zones noires). Cette perte d'informations peut avoir plusieurs origines. Des petites zones sont dues à la parallaxe entre le projecteur et le récepteur. Par contre de plus grandes zones sans informations sont sans doutes dues à la présence des rayons de soleil. Le capteur de profondeur est en effet sensible aux rayons

infrarouges émis dans la même bande spectrale que celle de son émetteur. Le capteur peut également être perturbé par des surfaces absorbantes (certains écrans d'ordinateur), par rayonnements infrarouge (non réfléchissantes) émis par des surfaces ou même par la perte de champs de vision pour son récepteur. Le fait de traiter ces pixels comme une information de profondeur normale entraîne un biais avec la majorité des méthodes d'apprentissage profond et dégrade la qualité des applications qui utilisent l'image de profondeur. En effets les valeurs nulles de pixels peuvent créer soit des faux positifs, soit des faux négatifs dans l'apprentissage.

Ainsi, il est nécessaire d'ajouter une étape de pré-traitement afin de remplir les trous dans

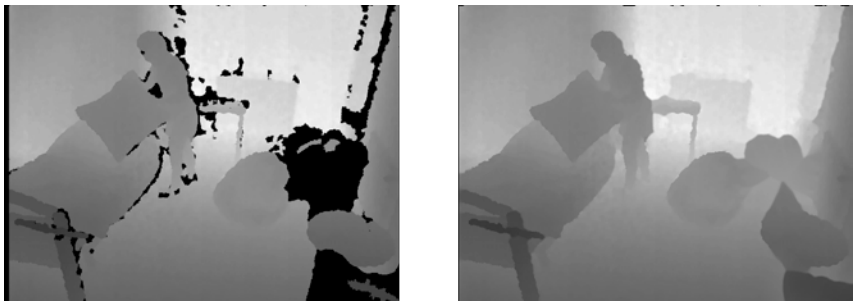


FIGURE 7.3 – Technique de remplissage des trous (Inpainting en anglais).

l'image (inpainting en anglais). L'inpainting est un domaine de recherche relativement nouveau qui présente des problématiques intéressantes pour les images de profondeur. Les méthodes de remplissage des trous au niveau des images couleurs sont performantes et bien maitrisées, mais les images de profondeur ont des caractéristiques totalement différentes qui peuvent dégrader les performances des techniques d'images couleurs. Des nouvelles techniques dédiées pour les images de profondeur ont été proposées afin de compenser la perte d'information de profondeur [161, 162, 163, 164, 165, 166, 167]. Les méthodes de remplissage de trous sur des images de profondeur ont été divisées en trois catégories : 1) les méthodes basées sur l'information spatiale [168, 169, 170], 2) les méthodes basées sur l'information temporelle [171, 172] et 3) la combinaison de ces deux informations [173, 174]. La première catégorie utilise les valeurs des pixels voisins disponibles dans une seule image pour compléter les données manquantes sur l'image en profondeur [175].

Nous avons repris cette catégorie de reconstruction de l'image de profondeur en se basant sur la technique proposée par Telea [161]. Il essaye d'estimer les valeurs correctes de ces pixels non définis à partir d'une moyenne pondérée sur leur voisinage non nul (Fig. 7.3). L'objectif de cette reconstruction n'est pas d'avoir une image la plus proche possible

de la réalité mais de minimiser au maximum les erreurs induites par les pixels nuls lors de l'apprentissage et classification des activités.

7. 1.2 Augmentation des données

Il est essentiel de disposer d'une large base de données avec des situations différentes pour la performance du modèle d'apprentissage et pour éviter le surapprentissage (*over fitting*). La technique d'augmentation des données (*Data augmentation*) est souvent utilisée pour améliorer les capacités de généralisation des modèles. Les modifications appliquées sur les données réelles sont aléatoires mais doivent rester réalistes, comme par exemple la rotation, l'ajout d'un bruit gaussien, la suppression de petites régions, l'effet miroir ou le zoom. Nous avons effectué ces modifications sur notre base d'apprentissage. La Figure 7.4 illustre un exemple de ces modifications.
L'idée de cette étape est d'appliquer de 1 à 3 modifications sur la même image. Le nombre et le choix de types de modification est aléatoire afin de diversifier les données existantes et ne pas avoir les mêmes modifications sur toute la base.

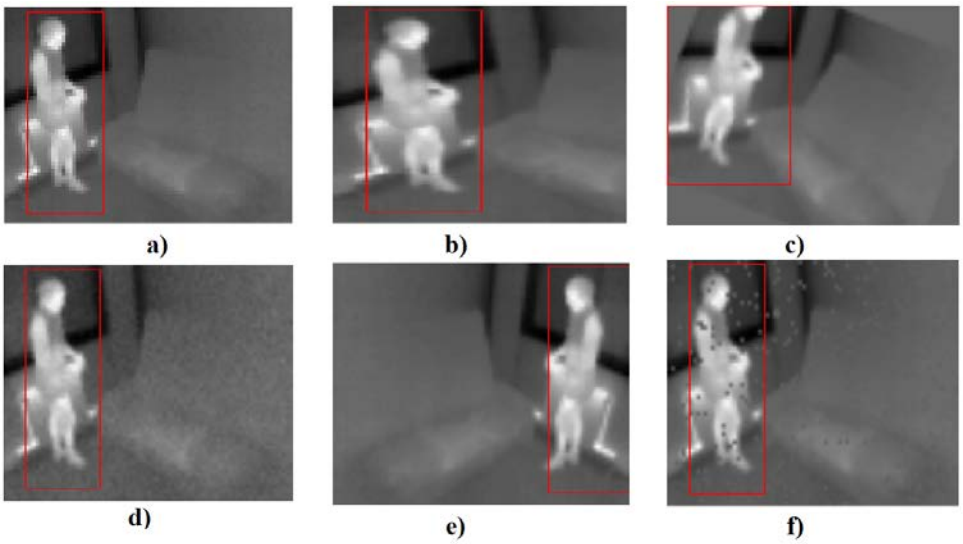


FIGURE 7.4 – Exemple d'augmentation de données sur l'image d'origine (a) : zoom (b), rotation (c), bruit gaussien (d), effet miroir (e) et suppression de régions (f).

7. 2 Méthode SSD

Les réseaux d'apprentissage profond sont de plus en plus performants pour la classification des images. Ils peuvent même dépasser les capacités humaines dans certains cas complexes [158]. SSD est un algorithme de classification des images et de détection des objets.

SSD, comme le montre la Figure 7.5, est construit autour de deux composants. Le premier composant, appelé réseau de base, est pré-entrainé pour extraire des caractéristiques de *classification des images*. Deux types de réseau convolutif sont généralement utilisés dans le *SSD* pour cette classification : le *VGG16* [176] ou le *ResNet* [177]. Dans notre

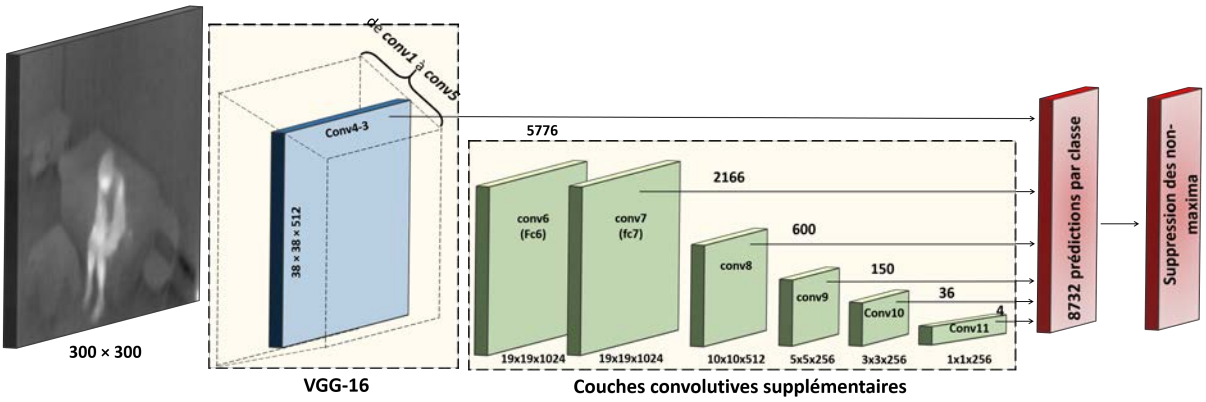


FIGURE 7.5 – L'architecture globale du réseau SSD.

projet, nous avons choisi d'utiliser le *VGG16*, jusqu'à la couche *conv-5*. Le deuxième composant est utilisé pour *détecter et localiser* les objets sur chaque image. Ensuite, une couche de *prédiction* fusionne des cartes de caractéristiques multi-échelles issues des couches convolutives pour générer des boîtes englobantes associées à des probabilités contenant les objets d'intérêt. Par la suite, une étape de *suppression des non-maxima* est utilisée pour garder les détections potentiellement correctes.

Nous allons détailler un peu le principe de la détection d'objet. Elle consiste à identifier et à localiser l'objet par une boîte englobante. Trois notions sont utilisées pour détecter cet objet : les cellules, les boîtes englobantes et le rapport hauteur/largeur.

Cellule, appelée également localisation (Cell). Au lieu d'utiliser une fenêtre coulissante, *SSD* divise l'image en utilisant une grille et fait en sorte que chaque cellule de la grille soit responsable de la détection des objets dans cette région de l'image. La taille de

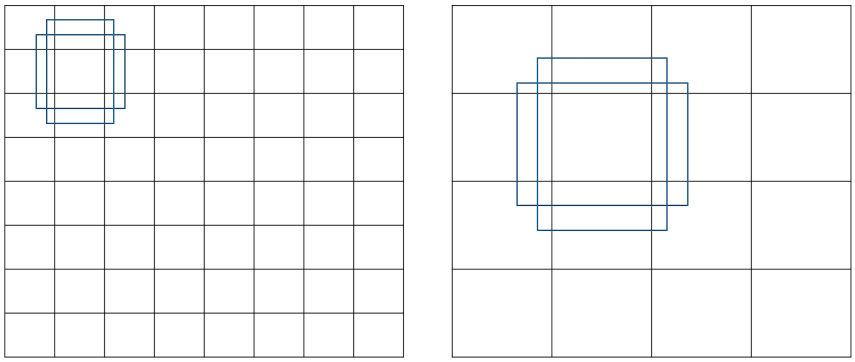


FIGURE 7.6 – Carte de référence à faible résolution (à droite), divisée en 4×4 cellules et carte de référence à moyenne résolution (à gauche), divisée en 8×8 cellules. Des boîtes englobantes sont ensuite associées à ces cellules (en bleu).

la grille est importante pour détecter les objets en fonction de leur échelle (Figure 7.6). Par exemple, nous pouvons utiliser une grille (4×4) pour détecter les objets à grande échelle ou des grilles plus fines (8×8) pour objets de plus petite échelle. Au final, chaque cellule de la grille est capable de fournir la position et la forme de l'objet qu'elle contient.

Boîtes englobantes (Anchor box). Á chaque cellule de la grille du *SSD* est associée plusieurs boîtes englobantes. Ces boîtes sont prédéfinies et chacune d'entre elles est définie par une taille et une forme uniques dans une cellule de la grille. *SSD* utilise une phase d'appariement pendant l'entraînement, pour faire correspondre la boîte englobante avec les boîtes de vérité terrain de chaque objet dans une image. La taille et le nombre des boîtes englobantes sont détaillés dans la prochaine section.

Rapport hauteur/largeur (Aspect ratio). Tous les objets ne sont pas de forme carrée. Certains sont plus longs et d'autres plus larges, à des degrés divers. L'architecture *SSD* permet de tenir compte de ces différences dans les rapports d'aspect des boîtes englobantes. Le paramètre "ratio" peut être utilisé pour spécifier les différents rapports d'aspect des boîtes englobantes associées à chaque cellule de la grille.

7. 2.1 Approche

La détection d'objets à partir du réseau *SSD* se compose de 2 traitements :

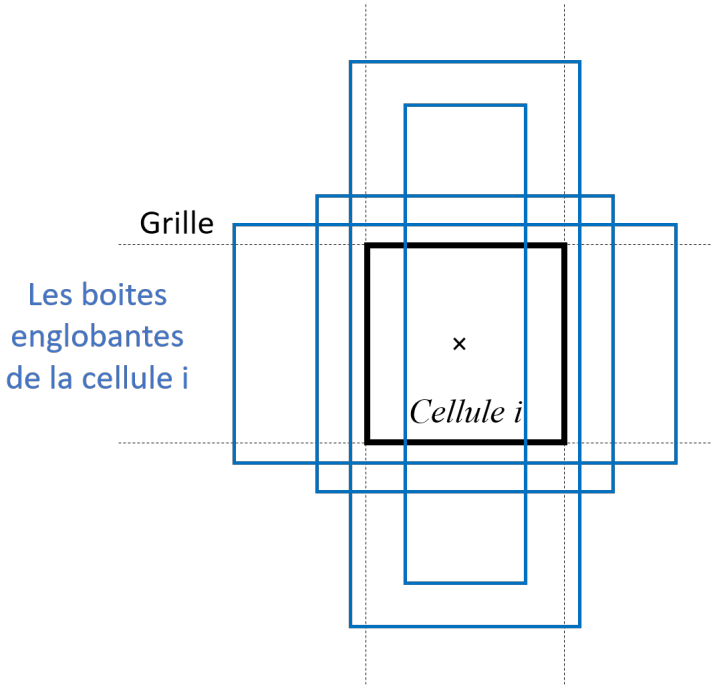


FIGURE 7.7 – Les différentes boîtes englobantes utilisées dans le SSD.

— **Extraire des cartes de caractéristiques.** *SSD* commence à détecter les objets à partir de la couche *Conv4-3* du réseau *VGG16*. Cette couche génère des cartes de caractéristiques 38×38. Pour chaque cellule, il y a quatre prédictions d'objets. Chaque prédiction se compose d'une boîte englobante de taille unique, comme le montre la Figure 7.7. Ainsi la couche *Conv4-3* fait un total de 5776 ($38 \times 38 \times 4$) prédictions : quatre prédictions par cellule quelque soit la profondeur des cartes de caractéristiques.

— **Appliquer des filtres de convolution pour détecter les objets.** Contrairement aux autres réseaux de classification, le *SSD* n'utilise pas le principe de deux étages pour localiser et classer les objets (pas de réseau de proposition des régions RPN). Il calcule à la fois les scores de localisation et de classification à partir des filtres de petites tailles. Après l'extraction des cartes de caractéristiques, chaque carte peut produire un ensemble fixe de prédictions en utilisant un ensemble de filtres convolutifs (Figure 7.5). Pour une carte de caractéristiques de taille $m \times n$ avec p canaux, des filtres de convolution de taille $(3 \times 3 \times p)$ sont appliqués soit pour générer un score d'appartenance à une classe, soit pour décaler la boîte à l'emplacement de l'objet.

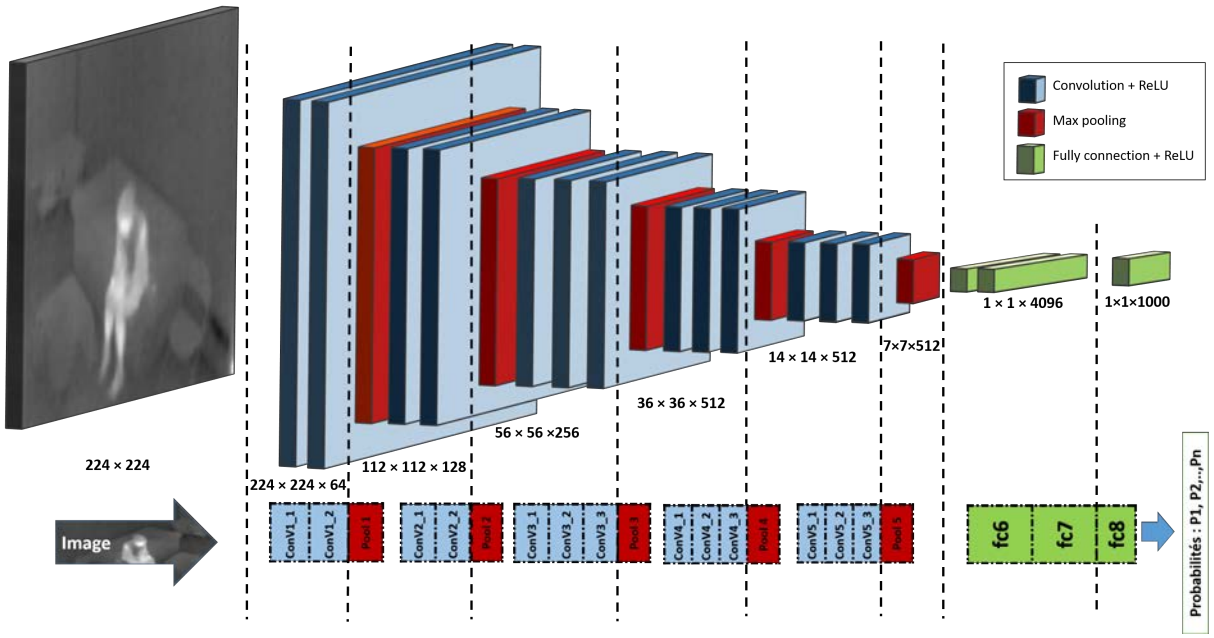


FIGURE 7.8 – L'architecture du réseau VGG16.

7. 2.2 Architecture du VGG16

L'architecture de VGG16 est représentée sur la Figure 7.8. L'entrée de la première couche *conv*1 est une image RGB de taille 224×224. L'image passe à travers une pile de couches convolutives, dont les filtres ont une taille (3×3).
La sortie de VGG utilisée dans le réseau SSD est la carte de caractéristiques de la couche *conv4-3* comme modèle de base pour l'extraction des caractéristiques utiles. En plus du *VGG16*, le réseau *SSD* contient d'autres couches de caractéristiques de tailles décroissantes. Ces couches peuvent être considérées comme une représentation pyramidale d'images à différentes échelles.

7. 2.3 Architecture du SSD

La Figure 7.5 illustre l'architecture du *SSD*. À partir de *conv5-3*, six couches convolutives ont été ajoutées. Chaque couche ajoutée permet de produire des prédictions. Ce qui mène à *8732* prédictions à la sortie du réseau *SSD*. La sortie de la couche *Conv4-3* ainsi que ceux des cinq dernières couches supplémentaires se connectent à la couche de détection finale. Cette connexion permet au réseau de détecter et de localiser des objets à différentes échelles. La Table 7.1 représente les prédictions de chaque carte de caracté-

Carte de références	Dimensions	Échelle préalable	Rapport hauteur/largeur	Nombre des boîtes englobantes	Nombre total des prédictions
conv4_3	38x38	0.1	1:1, 2:1, 1:2 + autres	4	38x38x4=5776
conv7	19x19	0.2	1:1, 2:1, 1:2, 3:1, 1:3 + autres	6	19x19x6=2166
conv8_2	10x10	0.375	1:1, 2:1, 1:2, 3:1, 1:3 + autres	6	10x10x6=600
conv9_2	5x5	0.55	1:1, 2:1, 1:2, 3:1, 1:3 + autres	6	5x5x6=150
conv10_2	3x3	0.725	1:1, 2:1, 1:2 + autres	4	3x3x4=36
conv11_2	1x1	0.9	1:1, 2:1, 1:2 + autres	4	1x1x4=4
Total	–	–	–	–	**8732 prédictions**

TABLE 7.1 – Répartition des prédictions sur le réseau SSD.

ristiques connectée à la couche finale.

Le réseau *SSD* prédit les scores des classes pour chaque carte de caractéristiques qui spécifient la présence d'une instance de classe dans chacune de ces boîtes. La Figure 7.9 montre comment le réseau "voit" une image donnée à travers ses cartes de caractéristiques.

Des travaux précédents [179, 157] ont montré que l'utilisation de cartes de caractéristiques des couches inférieures améliore la qualité de la segmentation, car les couches inférieures capturent des détails plus fins des objets d'entrée. Dans cette optique, Liu et al [158] ont utilisé ces couches ainsi que les autres couches supérieures pour la détection et la classification des objets. Pour traiter les différentes échelles d'objets, certaines méthodes [180] suggèrent de traiter l'image à différentes tailles et de combiner les résultats par la suite. En utilisant des cartes de caractéristiques de plusieurs tailles dans un seul réseau, nous pouvons imiter le même effet. Par exemple si nous avons m cartes de caractéristiques pour la prédiction des classes, l'échelle (scale) des boîtes englobantes pour chaque carte de caractéristiques est calculée comme suit :

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1), k \in [1, m] \tag{7.1}$$

avec s_{min} égal à 0.1 et s_{max} est 0.9, ce qui signifie que la couche la plus à gauche a

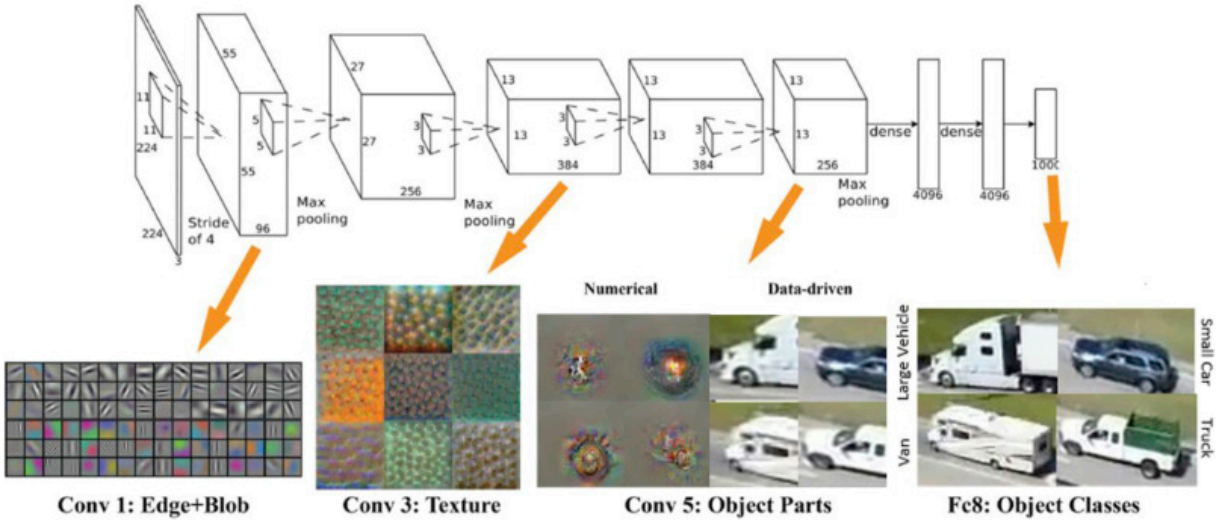


FIGURE 7.9 – Traitement du réseau à plusieurs étapes [178].

une échelle de $0,1$ et la couche la plus à droite a une échelle de $0,9$, et que toutes les couches intermédiaires sont régulièrement espacées. Différents rapports d'aspect (aspect ratio) pour les boîtes englobantes sont imposés par $a_r \in \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$, comme le montre la Table 7.1. Alors la largeur et la longueur de la boîte sont désignées respectivement par :

$$w_k^{a_r} = s_k\sqrt{a_r} \tag{7.2}$$

$$h_k^{a_r} = \frac{s_k}{\sqrt{a_r}} \tag{7.3}$$

Deux boîtes sont ajoutées, d'échelle :

$$s_k = \sqrt{s_k * s_{k+1}} \tag{7.4}$$

Ce qui donne six boîtes par emplacement pour les trois couches supplémentaires du réseau.

Nous détaillons dans la Table 7.2 l'architecture du réseau *SSD* en précisant les tailles des matrices à l'entrée de chaque couche ainsi que le nombre et la taille des filtres utilisés dans chaque convolution. Nous définissons aussi le pas de convolution, la valeur de "padding" et la taille des tenseurs [1] en sortie. La Table 7.2 récapitule le contenu de différentes parties du réseau *SSD*. Par exemple, les couches en jaunes sont les couches du *VGG16*.

1. C'est un tableau de dimensions $(n \times m \times p)$

Couches	Taille de matrice d'entrée			Taille de filtres	Nombre de filtres	Stride	Padding	Tenseur de sortie		
	H	W	Channels					H	W	Channels
Entrée	300	300	1							
Conv1_1	300	300	1	3	64	1	1	300	300	64
Conv1_2	300	300	64	3	64	1	1	300	300	64
Max_pool_1	300	300	64	2	64	2	0	150	150	64
Conv2_1	150	150	64	3	128	1	1	150	150	128
Conv2_2	150	150	128	3	128	1	1	150	150	128
Max_pool_2	150	150	128	2	128	2	0	75	75	128
Conv3_1	75	75	128	3	256	1	1	75	75	256
Conv3_2	75	75	256	3	256	1	1	75	75	256
Conv3_3	75	75	256	3	256	1	1	75	75	256
Max_pool_3	75	75	256	2	256	2	0	38	38	256
Conv4_1	38	38	256	3	512	1	1	38	38	512
Conv4_2	38	38	512	3	512	1	1	38	38	512
Conv4_3	38	38	512	3	512	1	1	**38**	**38**	**512**
Max_pool_4	38	38	512	2	512	2	0	19	19	512
Conv5_1	19	19	512	3	1024	1	1	19	19	1024
Conv5_2	19	19	1024	3	1024	1	1	19	19	1024
Conv5_3	19	19	1024	3	1024	1	1	19	19	1024
Conv6 (FC6)	19	19	1024	3	1024	1	1	19	19	1024
Conv7(FC7)	19	19	1024	1	1024	1	0	**19**	**19**	**1024**
Conv8_1	19	19	1024	1	256	1	0	19	19	256
Conv_8_2	19	19	256	3	512	2	1	10	10	**512**
Conv9_1	10	10	512	1	128	1	0	10	10	128
Conv9_2	10	10	128	3	256	2	1	**5**	**5**	**256**
Conv10_1	5	5	256	1	128	1	0	5	5	128
Conv_10_2	5	5	128	3	256	1	0	**3**	**3**	**256**
Conv11_1	3	3	256	1	128	1	0	3	3	128
Conv11_2	3	3	128	3	256	1	0	**1**	**1**	**256**

TABLE 7.2 – Architecture du réseau SSD.

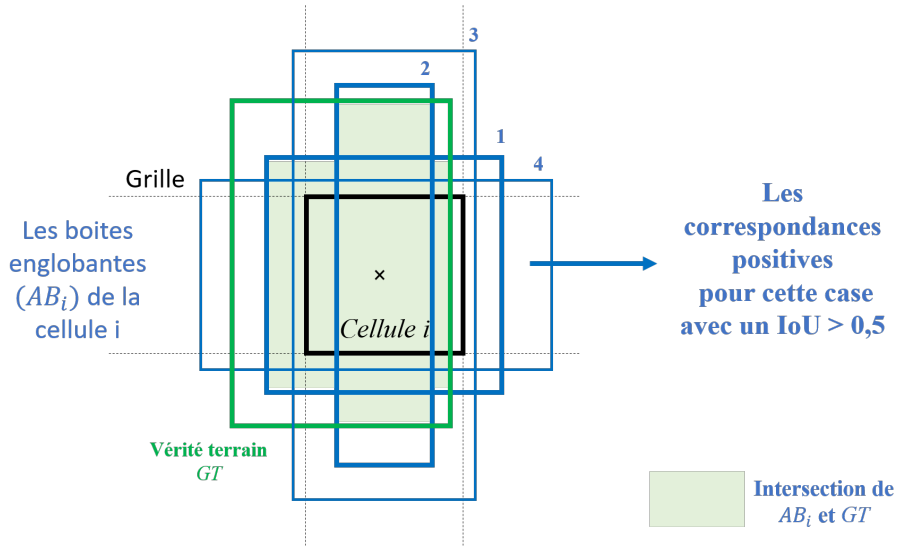


FIGURE 7.10 – Chevauchement des boîtes englobantes (en bleu) avec la vérité terrain (en vert) pour détecter les correspondances positives.

Les prédictions du *SSD* sont classées comme des correspondances positives ou négatives. L'aspect positif ou négatif de la correspondance est donnée par l'IoU, l'intersection sur l'union, qui est le rapport entre la zone d'intersection et la zone jointe pour deux régions (éq. 8). Si la boîte a un IoU supérieur à 0,5 avec la vérité de terrain, la correspondance est positive. Dans le cas contraire, elle est négative. Prenons l'exemple de la Figure 7.10, seules les boîtes 1 et 2 sont des correspondances positives. Le *SSD* n'utilise que les correspondances positives pour calculer le coût de localisation (le décalage de la boîte englobante).

Fonction de coût globale. L'objectif de l'apprentissage est d'entraîner le réseau à minimiser une fonction de coût globale qui mesure la différence entre la sortie du réseau et la vérité terrain. Soit $x_{ij}^p = \{0, 1\}$ un indicateur de correspondance entre la i-ème cellule par défaut et la j-ème boîte de vérité terrain de la classe p (éq. 7.5).

$$x_{ij}^p = \{^{1 \ si \ IoU>0.5}_{sinon \ 0} \tag{7.5}$$

La fonction de coût globale est une somme pondérée de la fonction de coût de localisation (L_{loc}) et de confiance (L_{loc}) (éq. 7.6) :

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \tag{7.6}$$

avec N le nombre de correspondances positives des boîtes englobantes (Dans le cas particulier où $N = 0$, $L = 0$) et α est le poids de pondération (dans [0,1]) de l'erreur de localisation.

La fonction de coût de localisation (L_{loc}) entre les paramètres de prédiction et de la boîte de vérité terrain ($l(l_x^p, l_y^p, l_w^p, l_h^p)$ *et* $g(g_x^p, g_y^p, g_w^p, g_h^p)$) respectivement, de la même classe p, est la fonction de coût *smooth L1 loss* [181] pour se décaler vers le centre c_x et c_y de la boîte par défaut d de largeur w et de hauteur h :

$$L_{Loc}(x, l, g) = \sum_{m \in x,y,w,h} x_{ij}^k \; smooth_{L1}(l_m^p - g_m^p) \tag{7.7}$$

$$avec \; smooth_{L1}(z) = \begin{cases} 0.5z^2 & si \;\; |z|<1 \\ sinon & |z|-0.5 \end{cases} \tag{7.8}$$

Concernant la classe "arrière-plan" ($p = 0$), elle ne possède pas des boîtes englobantes de vérité terrain. La fonction L_{loc} est donc éliminée. *smooth L1* peut être interprétée comme une combinaison de la fonction *L1 loss* et la fonction *L2 loss*. Elle se comporte comme la fonction L1 lorsque la valeur absolue de z est élevée, et comme la fonction L2 lorsque la valeur absolue de z est proche de zéro. *smooth L1* combine les avantages de L1 (gradients réguliers pour les grandes valeurs de z) et de la fonction L2 (moins d'oscillations lors des mises à jour lorsque z est petit). Elle est moins sensible aux valeurs aberrantes que la perte L2 utilisée dans R-CNN et SPPnet [181]. Lorsque les objectifs de régression ne sont pas limités, l'apprentissage avec la perte L2 peut nécessiter un réglage des taux d'apprentissage afin d'éviter l'explosion des gradients. L'équation 7.8 élimine cette sensibilité. L'erreur de confiance (L_{conf}) est l'erreur d'une prédiction de classe. Pour chaque prédiction de correspondance positive, nous pénalisons l'erreur en fonction du score de confiance de la classe correspondante [158]. Pour les prédictions de correspondance négatives, nous pénalisons l'erreur en fonction du score de confiance de la classe "0" : la classe "0" indique qu'aucun objet n'est détecté. Elle est calculée à partir

de la fonction *softmax* sur plusieurs scores de confiance des classes (c) :

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} log(\widehat{c_i^p}) - \sum_{i\in Neg} log(\widehat{c_i^0}) \tag{7.9}$$

avec $\widehat{c_i^p} = \frac{exp(c_i^p)}{\sum_p (c_i^p)}$.

8732 prédictions sont détectées par le réseau *SSD* pour avoir une meilleure détection de l'objet ainsi que sa localisation. *SSD* utilise donc plus de prédictions que de nombreuses autres méthodes de détection. Cependant, de nombreuses prédictions ne contiennent aucun objet. Par conséquent, toute prédiction dont le score de confiance de classe est inférieur à $0,01$ sera éliminée.

Suppression des non-maxima (non maximum suppression nms). Le *SSD* utilise une technique de suppression des non-maxima pour supprimer les prédictions en double sur le même objet. Le réseau trie les prédictions en fonction des scores de confiance (c). En partant de la prédiction de confiance la plus élevée, le réseau SSD évalue si des boîtes précédemment prédites ont un IoU supérieur à $0,45$ avec la prédiction actuelle pour la même classe. Si c'est le cas, la prédiction actuelle sera ignorée.

7. 3 Implémentation

7. 3.1 Base de données et matériel

Nous avons détaillé dans le deuxième chapitre (section 2.4.2) la phase de création de notre prope base de données. Nous avons fait les acquisitions dans trois locaux différents (un livingLab et deux appartements). Nous avons mis dans la base de test des images d'une base de données (GLADIS) créée dans une maison de retraite à Bruz. Cette dernière base n'est pas publique, elle a été créée précédemment par notre équipe [47]. Deux personnes très âgées ont participé à cette campagne d'enregistrement. La Table 7.3 contient la composition de chaque classe. Nous avons annoté 2600 images (thermique et profondeur) manuellement en encadrant à chaque fois la silhouette dans l'image. Pour chaque couple d'images annotées, un fichier de type *XML* est créé contenant les informations détaillées (emplacement, hauteur et largeur) sur les objets annotés. Nous rappelons que nous traitons des images thermiques et de profondeur dont la texture de l'arrière-plan est pauvre par rapport aux images couleurs.

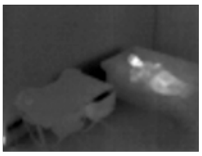
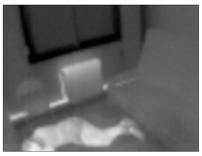
Classe	Assis (Sitting)	Debout (Standing)	Allongé sur le lit (Lying down)	Allongé par terre (Fall)	Total
					
Base d'apprentissage	160	246	163	257	824
Base de validation	30	67	39	72	207
Base de test	49	80	36	101	266
Total	239	393	238	430	1300

TABLE 7.3 – Base de données dédiée pour la reconnaissance des postures.

Nous avons exécuté les traitements sur une machine qui contient une carte graphique Nvidia RTX 2080 Ti, un processeur Intel Core i7-9700k, une mémoire RAM de 48 GB sous un système Linux (distribution Ubuntu).

7. 3.2 Résultats et discussion

Le modèle SSD a été appris à partir de 824 paires d'images et validé sur 207 paires d'images comme détaillées dans la Table 7.3, avec une taille de lot (batch size en anglais) de 16. Nous avons fixé le taux d'apprentissage initial (Learning Rate en anglais) à $LR = 0.0001$. La stratégie de sauvegarde du modèle de détection est basé sur l'hypothèse d'arrêt anticipé (Early Stopping en anglais), ce qui mène à sauvegarder le modèle dont l'erreur sur la base de validation (Loss validation) n'a pas été améliorée pendant 20 époques successives.

Résultats avec les images de profondeur. Nous présentons dans cette partie les résultats sur la seule base d'images de profondeur. La Figure 7.11 représente des résultats du modèle sur différentes images en classifiant les postures à chaque exemple.

Nous constatons sur ces exemples que les prédictions du modèle localisent la personne et détectent sa posture correctement. Pour évaluer cette méthode, nous avons calculé les métriques suivantes (Table 7.4) : la précision, le rappel, la moyenne de précision (AP) (éq. 10) pour chaque posture et la moyenne des AP (MAP) (éq. 11).

Pour analyser la performance de classification du modèle, nous avons schématisé les

FIGURE 7.11 – Résultats de classification des postures sur des images de profondeur : Standing = debout, sitting = assis, lying-down = allongé sur le lit et fall = allongé par terre.

	Précision	**Rappel**	**Précision moyenne (PA)**
Allongé par terre	0,9895	0,9793	0,905180
Allongé sur le lit	0,8750	0,9722	0,834726
Assis	0,9411	0,8889	0,791648
Debout	0,9746	0,9746	0,907862
mAP = 0,860			

TABLE 7.4 – Métriques d'évaluation sur des images de profondeur.

résultats obtenus dans une matrice de confusion (Figure 7.12.a). Le nombre dans la diagonale représente le nombre d'échantillons correctement prédits et la confusion du modèle pour chaque classe se trouve dans le reste des cases de la matrice. Nous avons normalisé les résultats de la matrice de confusion obtenus en fonction du nombre de données de tests (Figure 7.12.b). Nous avons constaté qu'en moyenne 96% de données ont été bien classées.

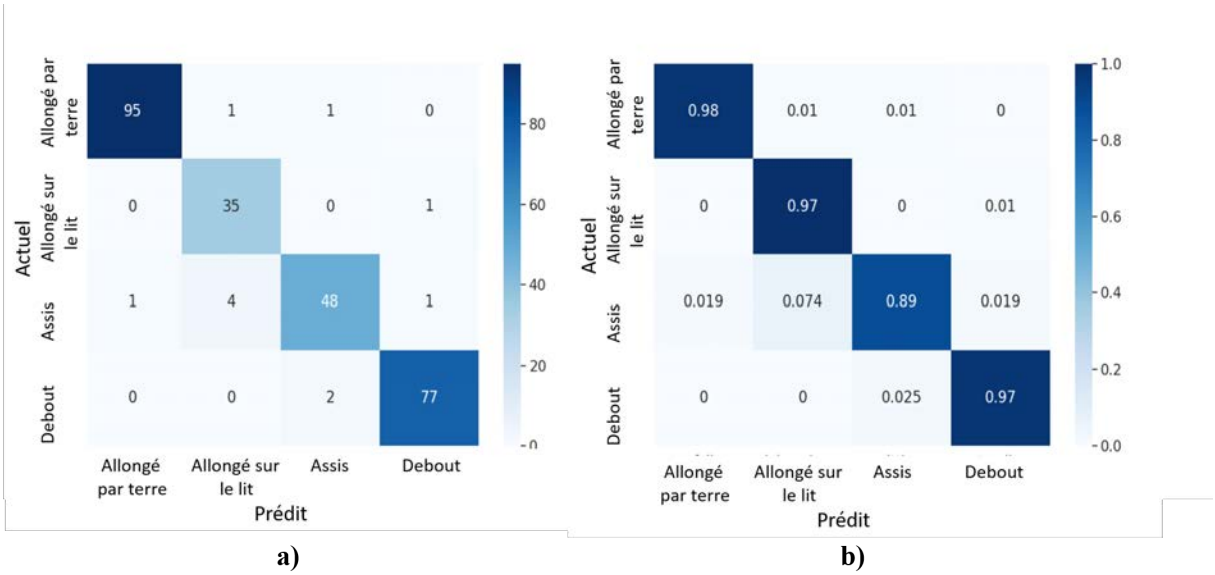


FIGURE 7.12 – Matrice de confusion des résultats de profondeur a) sans et b) avec normalisation.

Résultats des images thermiques. Nous présentons, dans cette section, notre système de reconnaissance de postures sur la base d'images thermiques contenant la même répartition que les données de profondeur. La Figure 7.13 représente des résultats du modèle sur différentes images en classifiant les postures à chaque exemple.

La Table 7.5 nous montre les mêmes métriques utilisées pour les images de profondeur afin d'évaluer le système de reconnaissance des postures sur les images thermiques.

	Précision	**Rappel**	**Précision moyenne (AP)**
Allongé par terre	0,9603	0,9700	0,908153
Allongé sur le lit	0,9722	0,9450	0,896232
Assis	0,9583	0,9013	0,886787
Debout	0,9382	0,9743	0,907910
mAP = 0,900			

TABLE 7.5 – Métriques d'évaluation sur des images thermiques.

La Figure 7.14.a illustre la matrice de confusion des résultats sur les images thermiques. La Figure (7.14.b) normalise les résultats obtenus selon la taille de la base de données et montre qu'en moyenne 95% des données ont été bien classées.

Comparaison entre les résultats thermiques et de profondeur. Afin de mieux comparer les résultats thermiques et de profondeur, nous avons également calculé le score F1 (éq. 6) qui combine la précision (éq. 1) et le rappel (éq. 4). Ce score est important pour estimer sur les bases, l'équilibre entre la précision et le rappel, car la précision seule ne peut pas identifier l'instabilité entre les classes. La précision, le rappel et le score F1 sont présentés dans la Table 7.6.

Nous constatons que les scores F1 de la posture "allongé par terre" sont élevés, environ $95,9\%$ en moyenne entre les deux types de bases de données. Par contre, la position assise semble être difficile à identifier dans certains cas, et est parfois confondue avec la posture debout ou allongée. La Figure 7.15 montre des exemples de détection sur des images thermiques et de profondeur. Dans cet exemple, nous observons que les détections sur les images de profondeur sont erronées, ce qui n'est pas le cas pour les images thermiques. Par

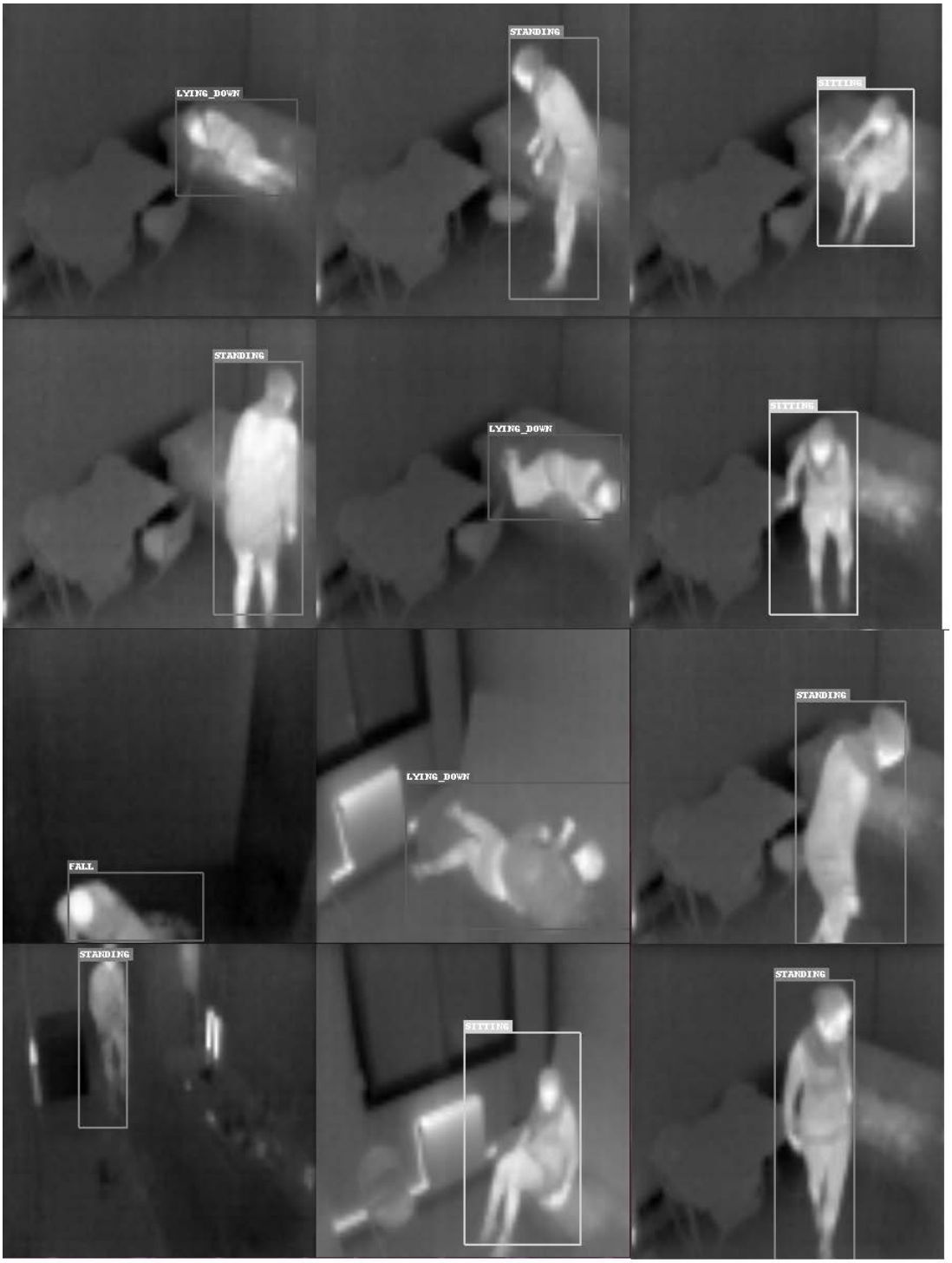


FIGURE 7.13 – Résultats de classification des postures sur des images thermiques.

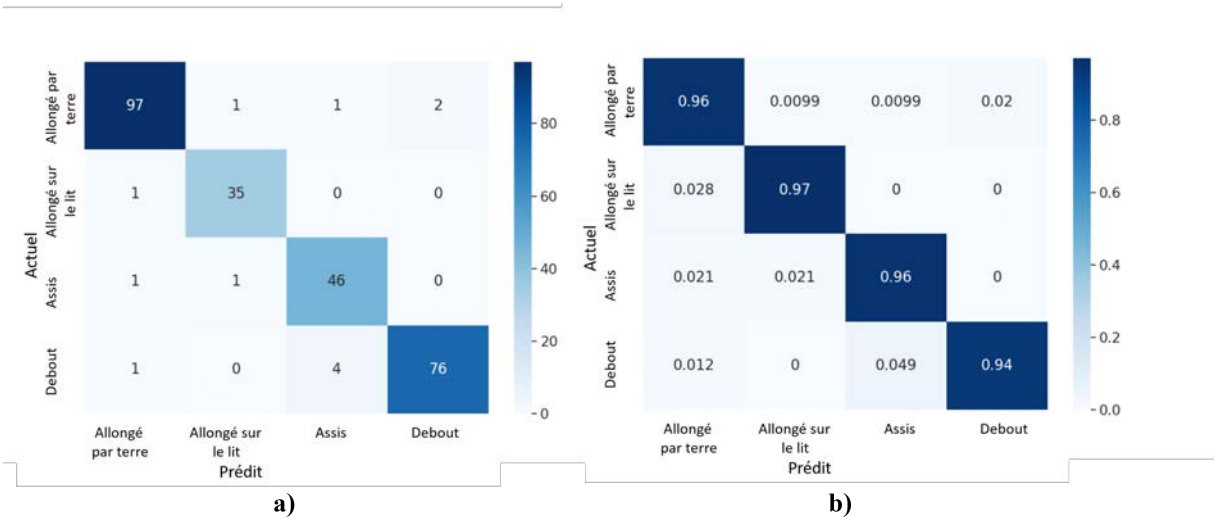


FIGURE 7.14 – Matrice de confusion des résultats thermiques sans et avec normalisation.

	Images de profondeur			Images thermiques		
Postures	**Précision**	**Rappel**	**Score F1**	**Précision**	**Rappel**	**Score F1**
Allongé par terre	0,9895	0,9793	**0,966948**	0,9603	0,9700	**0,965126**
Allongé sur le lit	0,8750	0,9722	0,953335	0,9722	0,9450	0,958407
Assis	0,9411	0,8889	0,918370	0,9583	0,9013	0,928926
Debout	0,9746	0,9746	0,962724	0,9382	0,9743	0,955909

TABLE 7.6 – Comparaisons des métriques obtenues sur les deux types d'images.

exemple, pour le premier couple d'images (à gauche), la personne essaie de se lever après une chute et est détectée "allongé par terre" sur l'image thermique mais "assis" sur l'image de profondeur. Nous rappelons que nous détectons les postures sur des images statiques où les transitions ne sont pas encore traitées. Prenons un autre exemple de détections où les erreurs sont repérées sur les images thermiques (Figure 7.16). Nous constatons, par exemple que sur le premier couple d'images (à gauche), la personne est en train de se coucher. Elle a été détectée "allongée sur le lit" sur l'image de profondeur mais "allongé par terre" sur l'image thermique. Donc dans certains cas, la bonne classification est donnée par l'image de profondeur et dans d'autres cas par l'image thermique.

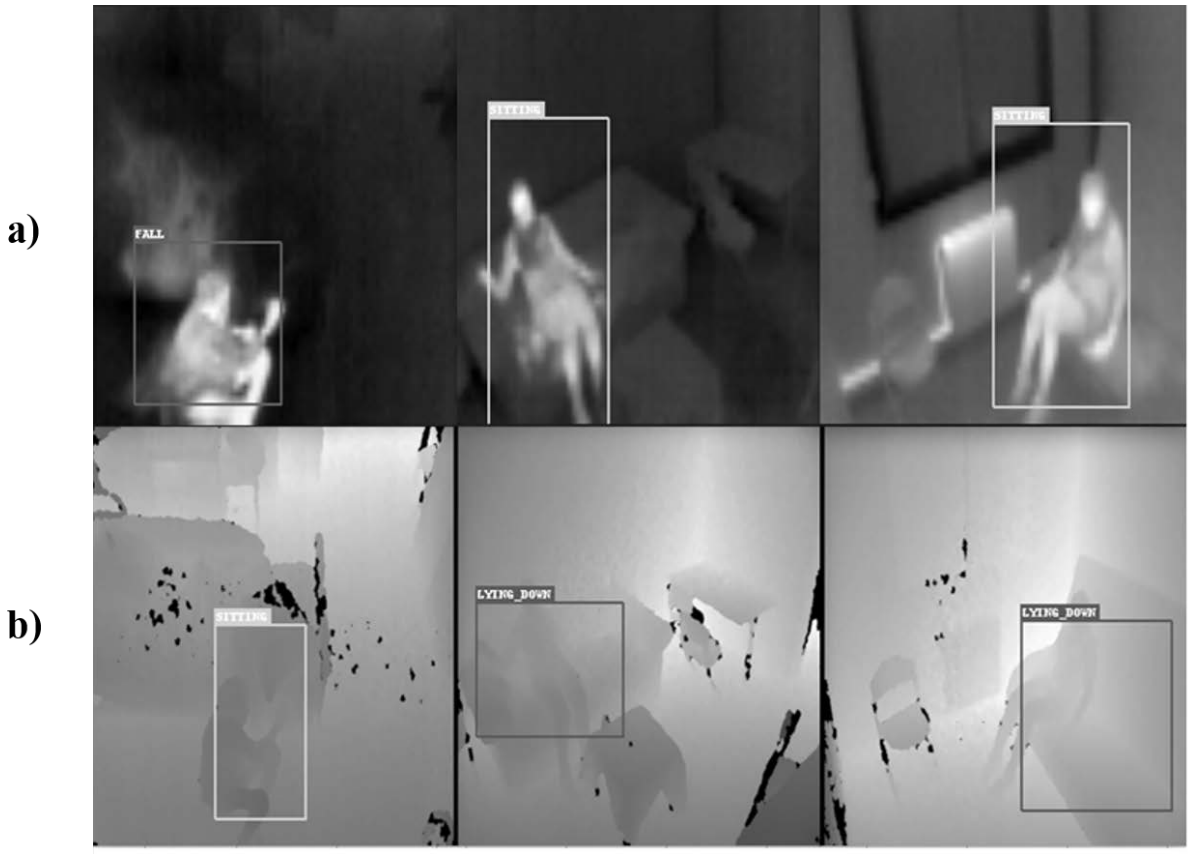


FIGURE 7.15 – Comparaison des résultats de détections sur des images thermiques (a) et des images de profondeur (b).

Pour améliorer ces résultats, nous avons décidé de fusionner les deux informations et une autre campagne d'acquisition est nécessaire.

7. 4 Fusion des décisions

7. 4.1 Principe

Nous avons constaté que les erreurs de détection ne sont pas identiques sur les images thermiques et de profondeur. Ainsi, nous avons décidé d'utiliser les deux informations en fusionnant les décisions de chaque réseau. La Figure 7.17 présente l'architecture de cette proposition. L'idée est de faire apprendre deux réseaux en parallèle ayant des entrées différentes et de fusionner par la suite les sorties de chaque réseau de neurones en gardant la probabilité la plus élevée pour chaque classe

FIGURE 7.16 – Comparaison des résultats de détections sur des images thermiques (a) et des images de profondeur (b).

Prenons l'exemple de trois scènes (A,B et C : chaque scène est associée à une image thermique et une image de profondeur) où nous avons deux détections différentes pour chaque type d'images, comme le montre la Table 7.7.

7. 4.2 Résultats

Nous avons créé deux réseaux SSD dont l'entrée de chacun est composé d'une image de taille (300×300). Puis, nous avons fusionné les décisions de chaque réseau en gardant la détection ayant la probabilité maximale. Nous avons détecté 20 fausses détections, dans la base de test sur la base d'images thermiques et 14 fausses détections pour la base de profondeur. En appliquant cette technique de fusion, nous avons rattrapé 10 détections par rapport au modèle thermique et 4 détections par rapport au modèle de profondeur, comme le montre la Table 7.8.

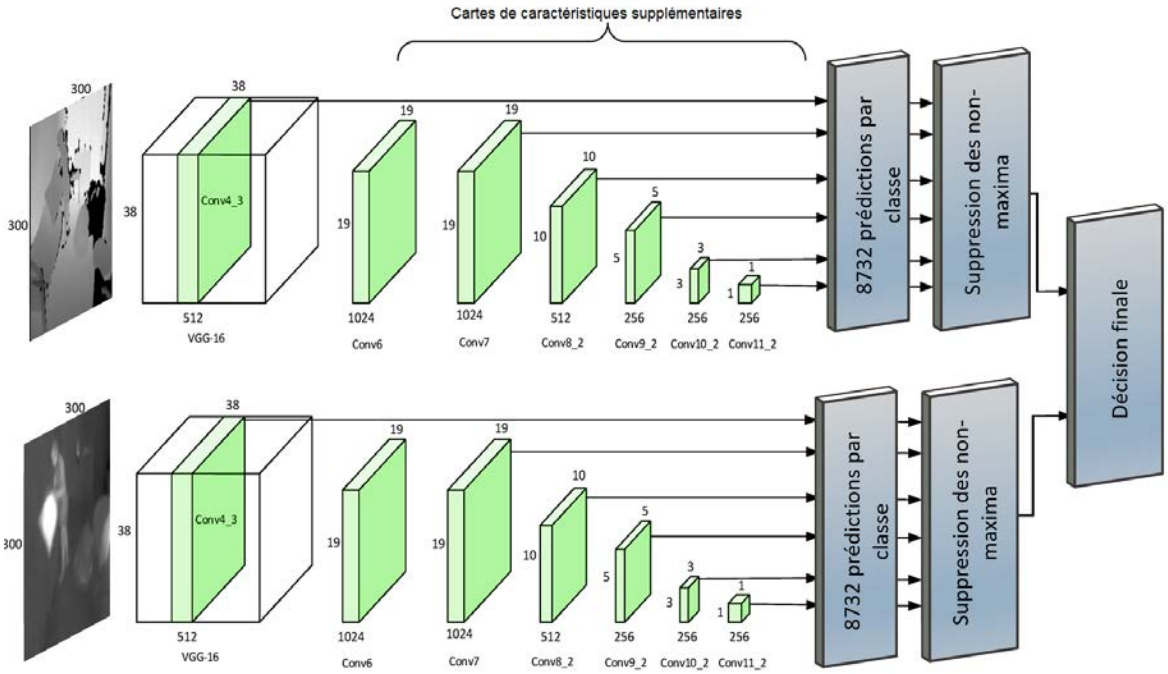


FIGURE 7.17 – Proposition de fusion de décisions avec les probabilités associées.

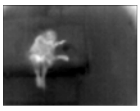
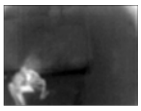
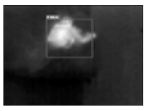
Fusion de décisions	**Image A**	**Image B**	**Image C**
Image thermique			
Vérité de terrain	assise	chute	chute
Détection thermique	chute (0,49)	assise (0,57)	chute (0,93)
Détection de profondeur	assise (0,90)	chute (0,83)	allongée (0,67)
Fusion = probabilité max	assise	chute	chute

TABLE 7.7 – Exemples de fusion de décisions.

La Figure 7.18 montre des exemples de classification de postures sur des images thermiques, de profondeur et de la fusion.

Pour évaluer ces résultats, nous avons utilisé les mêmes métriques des images thermiques et de profondeur. La Table 7.9 présente la précision, le rappel, la précision moyenne

	Thermique	Profondeur	Fusion
Fausses détections	20	14	10

TABLE 7.8 – Fausses détections avant et après fusion.

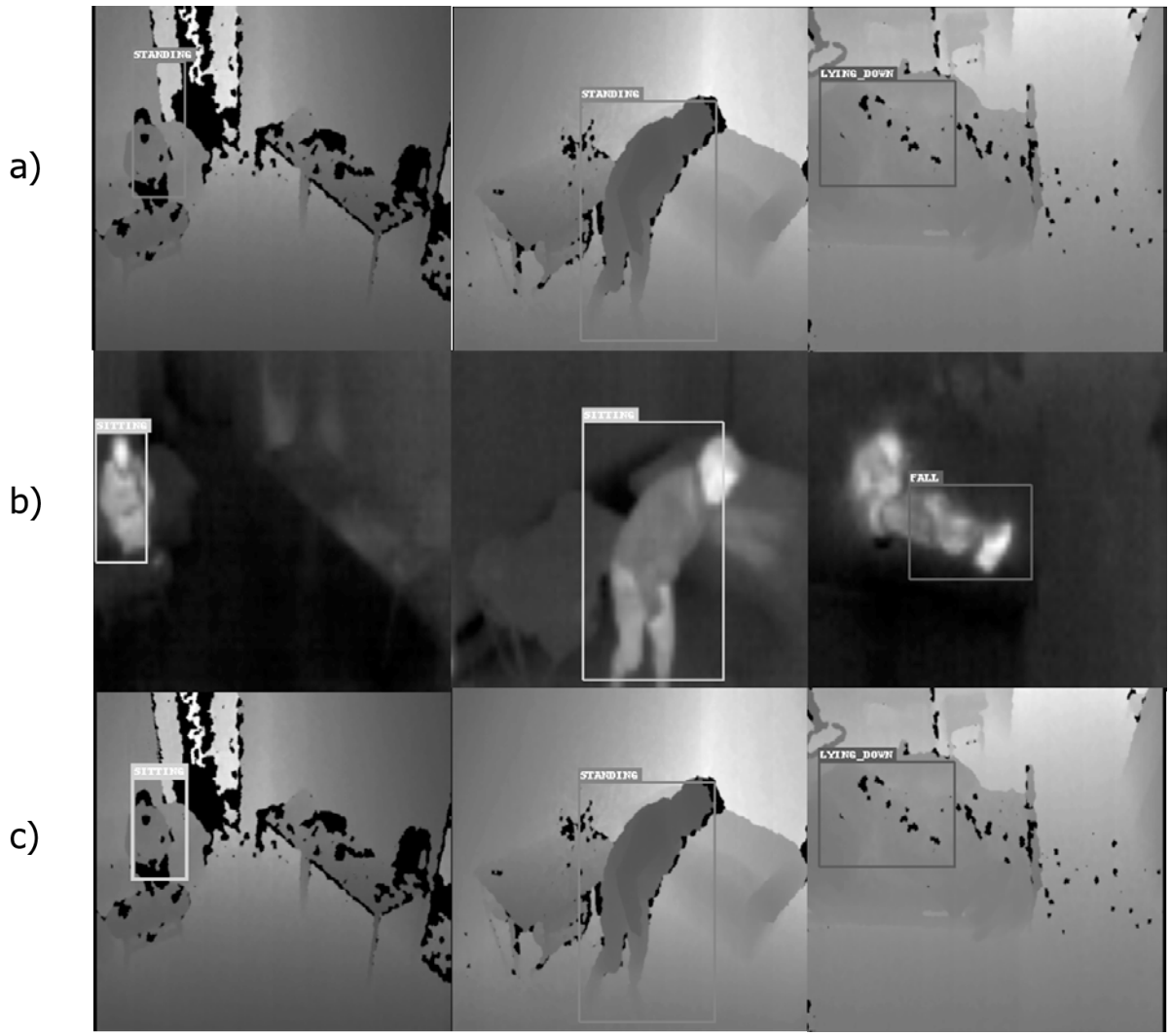


FIGURE 7.18 – Résultats de détection de postures en fonction a) du modèle de profondeur, b) du modèle thermique et c) du modèle de fusion.

(AP) et la mAP du modèle de fusion. Nous constatons que les précisions de différentes classes sont améliorées à part la classe "allongé par terre" et tous les rappels sont diminués,

ce qui mène à avoir une mAP inférieure aux autres obtenus précédemment.

	Précision	**Rappel**	**Précision moyenne (PA)**
Allongé par terre	0,98039	0,94339	0,9260
Allongé sur le lit	0,94736	0,97237	0,9230
Assis	0,91071	0,94444	0,8644
Debout	0,97530	0,98750	0,9635
mAP = 0,920			

TABLE 7.9 – Métriques d'évaluation de modèle de fusion.

Nous avons analysé ces résultats à l'aide de la matrice de confusion, comme la montre la Figure 7.19. Nous avons constaté que les vrais positifs de la posture "allongé par terre"

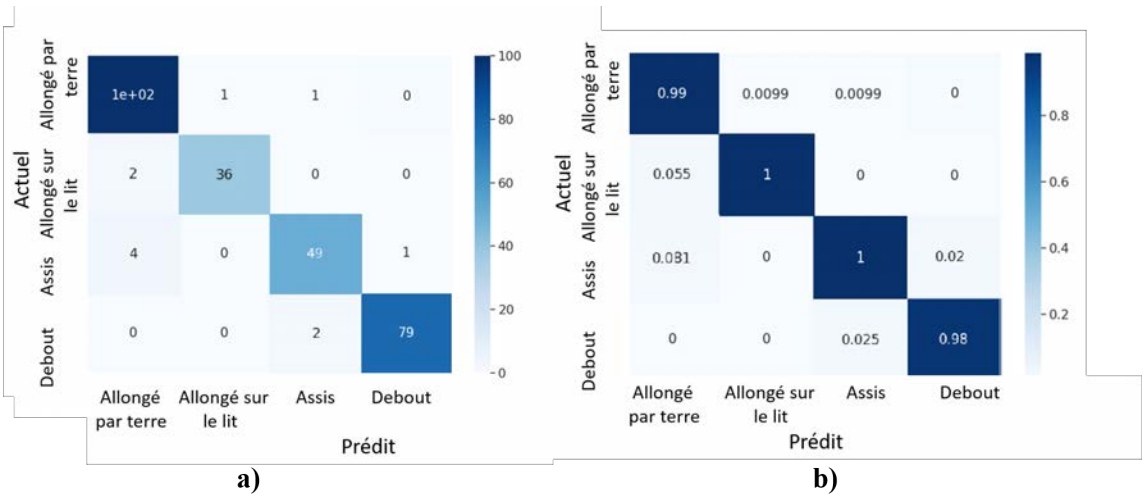


FIGURE 7.19 – Matrice de confusion de résultats de détection de postures en fonction du modèle de fusion.

ont augmenté, de même pour les faux négatifs, ce qui mène à avoir une baisse au niveau de la précision. Contrairement à la posture "allongé sur le lit", les vrais positifs et les faux négatifs ont diminué, ce qui aide à avoir un meilleur rappel mais une faible précision.

Pour comparer ce modèle avec la deuxième proposition, nous avons commencé par préparer la base de données des images fusionnées.

7. 5 Propositions de stratégie de fusion de données

Notre première stratégie de fusion décrites dans la section précédentes consistait à fusionner les décision des réseaux appliqués en parallèle sur les deux type d'images. Dans cette stratégie les deux réseaux sont appris de manière séparée, se privant ainsi de la complémentarité de l'information durant cette phase d'apprentissage. Il nous parait important de trouver des stratégies de fusion où soit les deux réseaux séparés communiquent entre eux (Figure 7.20) ou soit les deux images sont analysées ensemble par un même réseau (Figure 7.21).

Nous avons commencer par étudier cette seconde stratégie de fusionner les images qui consiste à donner en entrée du réseau une image vectorielle (une composante profondeur et une composante thermique) issue de la fusion des deux images

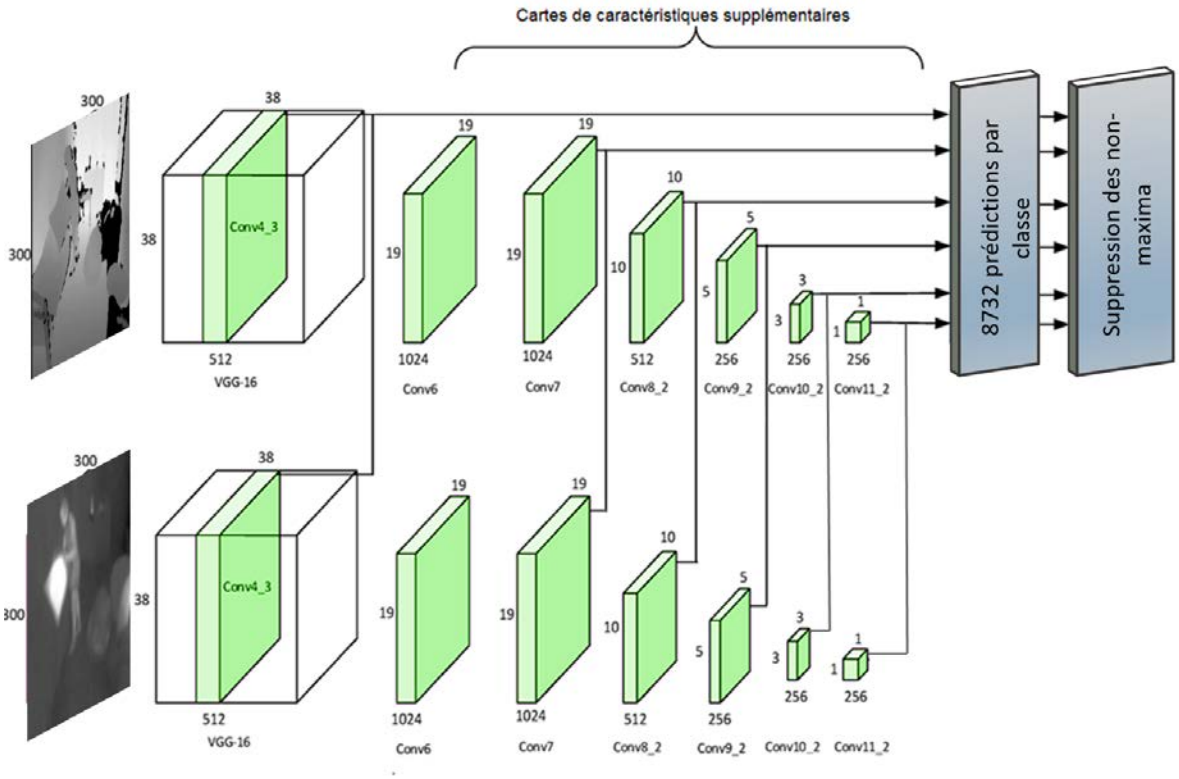


FIGURE 7.20 – Proposition de fusion en communiquant deux réseaux séparés.

Fusion des images. La Figure 7.22 illustre les différentes étapes qui nous permettent de fusionner les images de profondeur et thermique. Après la correction de déformations des

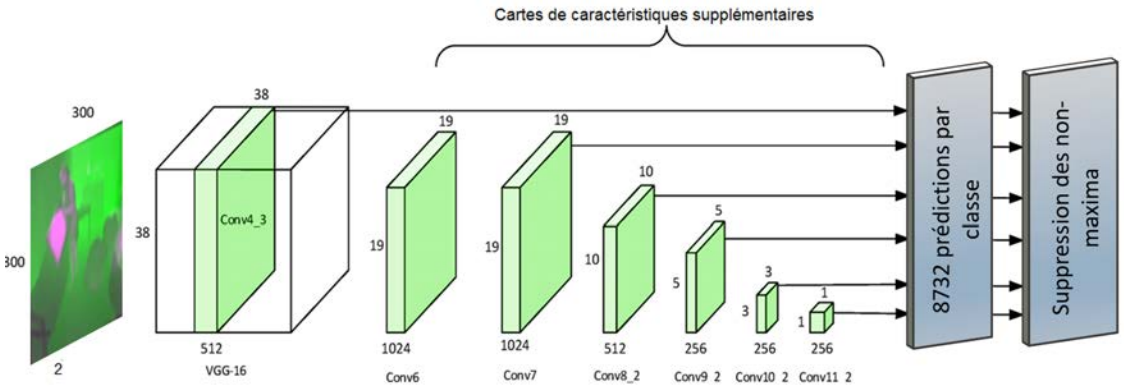


FIGURE 7.21 – Proposition de fusion au niveau de l'image d'entrée.

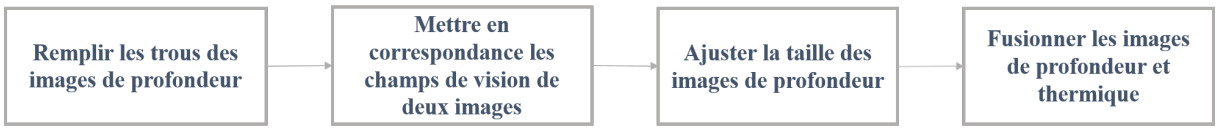


FIGURE 7.22 – Procédure de fusion des images de profondeur et thermique.

images de profondeur détaillée dans la section 7.1.1, une étape de mise en correspondance du champ de vision de l'image de profondeur avec celui de l'image thermique est réalisée. Cette étape est basée sur les paramètres de transformation de l'image de profondeur en image thermique obtenus durant la phase de calibration, détaillée dans la section 2.3. Le résultat final permet d'obtenir une image de profondeur de même champ visuel que celui de l'image thermique. La Figure 7.23 montre un exemple de cette transformation. Finalement, une dernière étape de redimensionnement des données est appliquée avant la superposition de deux images. Dans cette étape, l'image de profondeur est réduite pour avoir la même taille que celle de l'image thermique. Ce choix est justifié par rapport au temps de calcul. Nous avons utilisé une interpolation bilinéaire pour réduire la taille de l'image de profondeur. La Figure 7.24 montre le résultat final de l'étape de fusion utilisée. Faute de temps, nous n'avons réussi à évaluer que le premier modèle et le comparer avec les résultats de profondeur et thermiques obtenus auparavant. Nous envisageons de tester et évaluer les deux autres stratégies.

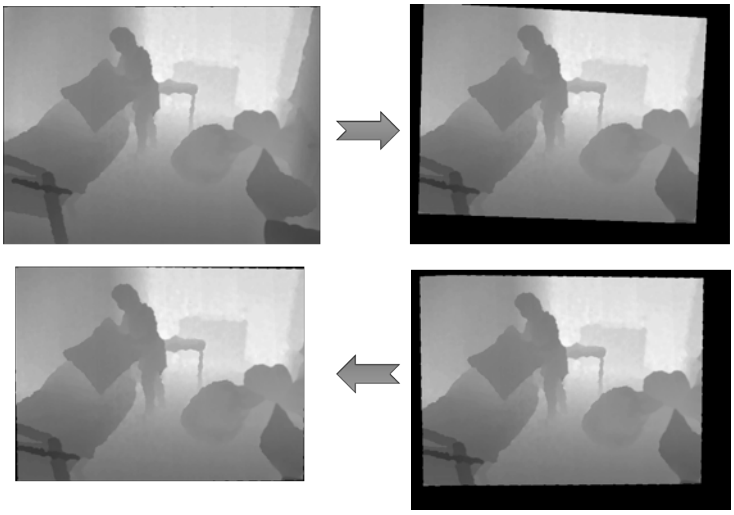


FIGURE 7.23 – Traitement de mise à l'échelle l'image de profondeur.

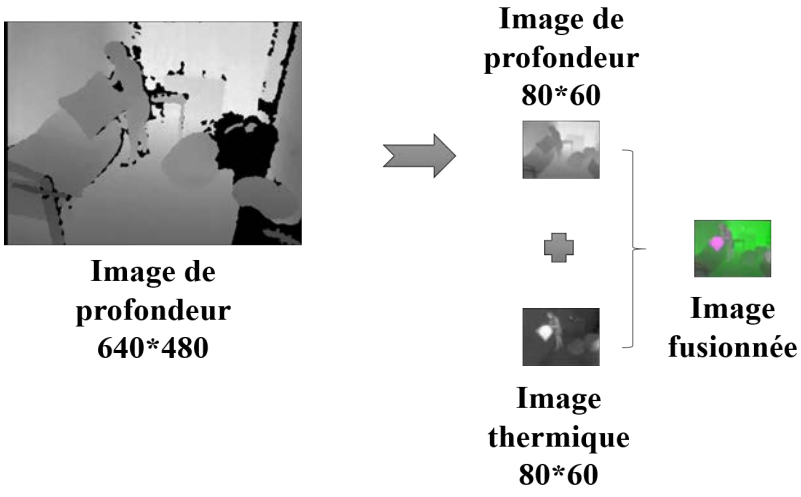


FIGURE 7.24 – Image multidimensionnelle issue de la fusion d'une image thermique et une image de profondeur modifiée.

7. 6 Conclusion

Dans ce chapitre, nous avons présenté une méthode de classification et localisation des postures de la personne âgée afin d'améliorer la détection des chutes à court terme et la prévention de la fragilité de la personne âgée à long terme. Cette méthode permet de distinguer quatre postures différentes (debout, assis, allongé sur le lit et au sol) en

appliquant une méthode d'apprentissage profond supervisé basée sur le réseau SSD (Single Shot Detector) sur les images thermiques et de profondeur. L'expérimentation de cette méthode a été réalisée suite à la création de la base de données de 824 couples d'images pour la phase d'apprentissage, 207 couples d'images pour la validation et 266 couples d'images pour la phase de test.

Nous avons remarqué quelques erreurs de classifications évitables. Nous avons donc décidé d'appliquer le principe de fusion pour améliorer nos résultats. Nous constatons que les détections ont été optimisées mais nous avons décidé de pousser les recherches encore plus loin et comparer cette version de fusion avec deux autres stratégies : 1) fusion de données et 2) fusion des réseaux qui restent à approfondir et tester.

Conclusion

Dans ce manuscrit, nous avons proposé un système de détection des chutes chez les personnes âgées en explorant deux stratégies différentes : une en suivant la personne lors de ses déplacements à l'intérieur de son logement et une autre en localisant la personne et en classifiant les postures de cette personne. Nous avons développé ce système de détection en se basant sur un dispositif automatique, à bas coût, qui fonctionne de jour comme de nuit et qui permet de préserver l'anonymat de la personne. Ce système, basé sur des capteurs de profondeur et thermique, est dédié aux personnes âgées autonomes résidant à leur propre domicile ou en EHPAD.

Dans un premier temps, nous avons détaillé les causes et les conséquences d'une chute chez une personne âgée ainsi que les différents systèmes de détection de chute existants jusqu'à présent. Nous avons classé ces systèmes en trois catégories selon le type de capteur utilisé pour analyser le comportement de la personne âgée. Ainsi, nous avons choisi les systèmes de vision parce qu'ils sont moins affectés par le bruit contrairement aux capteurs sonores. Ces systèmes se fixent une seule fois dans la pièce à surveiller et ils peuvent aussi être configurés une seule fois à l'installation, ce qui n'est pas le cas pour les capteurs portables. Concernant la protection de la vie privée, ces systèmes fournissent un moyen discret et non intrusif d'observation de la personne et de ses activités à l'exclusion des caméras couleurs.

Dans un second temps, nous avons présenté une méthode de suivi qui se base sur le filtrage particulaire dans un contexte de détection de chutes des personnes âgées. Pour commencer, nous avons segmenté la silhouette au niveau de l'image du premier plan, puis nous avons extrait l'ellipse qui englobe la tête. La représentation de la personne s'appuie sur la position, la taille et l'orientation de la tête. Ensuite, nous avons appliqué le filtrage sur l'ellipse extraite en utilisant la modalité de profondeur uniquement puis nous l'avons comparé à la segmentation seule. Enfin nous avons détaillé notre proposition de fusionner les informations de profondeur et thermiques en comparant deux modèles différents : un statique et un dynamique. Nous nous sommes intéressés au modèle dynamique qui met à jour les pondérations des informations de profondeur et thermiques, à chaque instant, selon l'importance de l'évolution de chaque information. De plus, nous avons ajouté la

vitesse de la tête comme nouveau paramètre pour la représentation de la personne. Ainsi, nous avons réussi à améliorer les performances du système pour atteindre les aires sous les courbes de précision et de succès égaux à 96.1% et 81.2%, respectivement.

Finalement, nous avons présenté une méthode de reconnaissance de la posture de la personne âgée afin d'améliorer la détection des chutes à court terme et la prévention de la fragilité de la personne âgée à long terme. Cette méthode permet de distinguer quatre postures différentes (debout, assis, allongé sur le lit et au sol) en appliquant une méthode d'apprentissage profond supervisé basée sur le réseau SSD (*Single Shot Detector*) sur les images thermiques et de profondeur. Cette partie a été réalisée suite à la création de notre propre base de données. Cette base est composée de 824 couples d'images pour la phase d'apprentissage, 207 couples d'images pour la validation et 266 couples d'images pour la phase de test. Ces images contiennent des exemples de situations dont l'arrière-plan n'a pas été appris par le réseau SSD. Nous avons atteint une précision moyenne (mAP) de toutes les postures de 90% pour les images thermiques et de 86% pour les images de profondeur. Nous avons identifié quelques détections dans lesquelles la classification n'est pas la même pour les deux types d'images (une correcte et une erronée). Nous avons donc décidé de fusionner les décisions des réseaux. Nous avons réussi à améliorer la précision moyenne (mAP) pour atteindre une valeur de 92%.

Pour conclure, nous avons participé au développement d'un système de suivi de la personne et de classification de ses postures sur les images thermiques et de profondeur. Ce système est relativement fiable pour analyser les activités de la personne âgée. Cependant, les perspectives de recherches et développements suivantes pourraient améliorer les performances de ce système :

- Fusionner les informations de profondeur et thermiques en entrée de réseau utilisé au niveau de la méthode de reconnaissance de postures afin d'enrichir la qualité des données d'entrée pour améliorer les performances de la méthode de classification.
- Tester un réseau de type LSTM pour analyser non pas des images statiques mais des séquences d'images.
- Appliquer une méthode de super-résolution sur les images thermiques pour améliorer la qualité de ces images et augmenter la précision de détection.
- Combiner le suivi de la personne et la reconnaissance de postures pour avoir une solution basée à la fois sur les aspects temporel et spatial, afin d'évaluer le comportement de la personne et de détecter les signes de sa fragilité à long terme.

BIBLIOGRAPHIE

[1] *Vieillissement et qualité de la vie,Les chutes*, https://www.who.int/ageing/about/facts/fr/ ,https ://www.who.int/fr/news-room/fact-sheets/detail/falls, Accessed : 2018-09-02.

[2] *DOSSIER : DÉTECTION AUTOMATIQUE DE CHUTE À DOMICILE*, http://www.hacavie.com/aides-techniques/essais-d-aides-techniques/articles/dossier-detection-automatique-de-chute-a-domicile/, Accessed : 2018-09-02.

[3] *La personne âgée et le risque de chute*, https://www.capretraite.fr/prevenir-dependance/perte-d-autonomie-et-maintien-a-domicile/la-personne-agee-et-le-risque-de-chute/, Accessed : 2019-03-16.

[4] Mozaffari NASSIM et al., « Practical Fall Detection Based on IoT Technologies : A Survey », in : *Internet of Things* 8 (oct. 2019), p. 100124, DOI : 10.1016/j.iot.2019.100124.

[5] Muhammad MUBASHIR, Ling SHAO et Luke SEED, « A Survey on Fall Detection : Principles and Approaches », in : *Neurocomputing* 100 (jan. 2013), p. 144-152, DOI : 10.1016/j.neucom.2011.09.037.

[6] Raúl IGUAL, Carlos MEDRANO et Inmaculada PLAZA, « Challenges, Issues and Trends in Fall Detection Systems », in : *Biomedical engineering online* 12 (juil. 2013), p. 66, DOI : 10.1186/1475-925X-12-66.

[7] James PERRY et al., « Survey and Evaluation of Real Time Fall Detection Approaches », in : *Proceedings of the 6th International Symposium High Capacity Optical Networks and Enabling Technologies*, jan. 2010, p. 158-164, DOI : 10.1109/HONET.2009.5423081.

[8] Aina RANDRIANARISAINA, Olivier PASQUIER et Pascal CHARGÉ, « Energy consumption modeling of smart nodes with a function approach », in : *Conference on Design and Architectures for Signal and Image Processing*, t. 2015, oct. 2014, DOI : 10.1109/DASIP.2014.7115645.

[9] Jai Sree, Uma A et Jenita B, « A survey on technical approaches in fall detection system », in : *National Journal of Physiology, Pharmacy and Pharmacology* 5 (sept. 2015), p. 1, doi : 10.5455/njppp.2015.5.0506201550.

[10] Kanitthika Kaewkannate et Soochan Kim, « A comparison of wearable fitness devices », in : *BMC Public Health* 16 (déc. 2016), doi : 10.1186/s12889-016-3059-0.

[11] Popescu Mihail et al., « An Acoustic Fall Detector System that Uses Sound Height Information to Reduce the False Alarm Rate », in : *International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, t. 2008, fév. 2008, p. 4628-31, doi : 10.1109/IEMBS.2008.4650244.

[12] Qiang Li et al., « Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information », in : *6th International Workshop on Wearable and Implantable Body Sensor Networks, BSN 2009*, juin 2009, p. 138-143, doi : 10.1109/BSN.2009.46.

[13] Majd Alwan et al., « A Smart and Passive Floor-Vibration Based Fall Detector for Elderly », in : *2nd International Conference on Information and Communication Technologies*, t. 1, jan. 2006, p. 1003-1007, doi : 10.1109/ICTTA.2006.1684511.

[14] Arni Ariani, David Chang et Nigel Lovell, « Simulated Unobtrusive Falls Detection With Multiple Persons », in : *IEEE transactions on bio-medical engineering* 59 (juil. 2012), p. 3185-96, doi : 10.1109/TBME.2012.2209645.

[15] Huan-Wen Tzeng, Mei-Yung Chen et Jai-Yu Chen, « Design of fall detection system with floor pressure and infrared image », in : *2010 International Conference on System Science and Engineering*, août 2010, p. 131-135, doi : 10.1109/ICSSE.2010.5551751.

[16] Nolwenn Lapierre et al., « The state of knowledge on technologies and their use for fall detection : A scoping review », in : *International Journal of Medical Informatics* 111 (déc. 2017), doi : 10.1016/j.ijmedinf.2017.12.015.

[17] Rita Cucchiara, Andrea Prati et Roberto Vezzani, « An Intelligent Surveillance System for Dangerous Situation Detection in Home Environments. », in : *Intelligenza Artificiale* 1 (jan. 2004), p. 11-15.

[18] Nicolas Thome, Serge Miguet et Sebastien Ambellouis, « A real-time, multi-view fall detection system : A LHMM-based approach », in : *Circuits and Systems for Video Technology, IEEE Transactions on* 18 (déc. 2008), p. 1522-1532, doi : 10.1109/TCSVT.2008.2005606.

[19] Homa Foroughi, Alireza Rezvanian et Amirhossien Paziraee, « Robust Fall Detection Using Human Shape and Multi-class Support Vector Machine », in : *Proceedings of the Sixth Indian Conference on CVGIP*, déc. 2008, p. 413-420, doi : 10.1109/ICVGIP.2008.49.

[20] Choon Lee et Alwyn Lee, « Fall Detection System Based on Kinect Sensor Using Novel Detection and Posture Recognition Algorithm », in : *International Conference on Smart Homes and Health Telematics*, t. 7910, juin 2013, p. 238-244, doi : 10.1007/978-3-642-39470-6_30.

[21] Thanh-Hai Tran, Thi Le et Jeremy Morel, « An analysis on human fall detection using skeleton from Microsoft kinect », in : *5th International IEEE Conference on Communications and Electronics, IEEE ICCE 2014* (oct. 2014), p. 484-489, doi : 10.1109/CCE.2014.6916752.

[22] J.W. Davis et V. Sharma, « Robust detection of people in thermal imagery », in : *International Conference on Pattern Recognition*, t. 4, jan. 2004, 713-716 Vol.4, isbn : 0-7695-2128-2, doi : 10.1109/ICPR.2004.1333872.

[23] Joseph Rafferty et al., « Fall Detection Through Thermal Vision Sensing », in : *International Conference on Ubiquitous Computing and Ambient Intelligence*, déc. 2016, p. 84-90, isbn : 978-3-319-48798-4, doi : 10.1007/978-3-319-48799-1_10.

[24] Javier Quero et al., « Detection of Falls from Non-Invasive Thermal Vision Sensors Using Convolutional Neural Networks », in : *International Conference on Ubiquitous Computing and Ambient Intelligence* 2 (oct. 2018), p. 1236, doi : 10.3390/proceedings2191236.

[25] Jacob Nogas, Shehroz Khan et Alex Mihailidis, « DeepFall : Non-Invasive Fall Detection with Deep Spatio-Temporal Convolutional Autoencoders », in : *Journal of Healthcare Informatics Research* 4 (déc. 2019), doi : 10.1007/s41666-019-00061-4.

[26] Jacob Nogas, Shehroz Khan et Alex Mihailidis, « Fall Detection from Thermal Camera Using Convolutional LSTM Autoencoder », in : *2nd Workshop on AI for Aging, Rehabilitation and Independent Assisted Living at IJCAI*, juil. 2018.

[27] Somasundaram Vadivelu et al., « Thermal Imaging Based Elderly Fall Detection », in : mar. 2017, p. 541-553, isbn : 978-3-319-54525-7, doi : 10.1007/978-3-319-54526-4_40.

[28] Lingmei Ren et Yanjun Peng, « Research of Fall Detection and Fall Prevention Technologies : A Systematic Review », in : *IEEE Access, vol. 7* (juin 2019), p. 1-1, doi : 10.1109/ACCESS.2019.2922708.

[29] Kabalan Chaccour et al., « From Fall Detection to Fall Prevention : A Generic Classification of Fall-Related Systems », in : *IEEE Sensors Journal* (nov. 2016), p. 1-1, doi : 10.1109/JSEN.2016.2628099.

[30] Xinyao hu et Xingda Qu, « Pre-impact fall detection », in : *BioMedical Engineering OnLine* 15 (déc. 2016), doi : 10.1186/s12938-016-0194-x.

[31] Nienke Kosse et al., « Sensor technologies aiming at fall prevention in institutionalized old adults : A synthesis of current knowledge », in : *International Journal of Medical Informatics* (juil. 2013), doi : 10.1016/j.ijmedinf.2013.06.001.

[32] Oladele Atoyebi, Antony Stewart et June Sampson, « Use of Information Technology for Falls Detection and Prevention in the Elderly », in : *Ageing International* 40 (sept. 2014), doi : 10.1007/s12126-014-9204-0.

[33] Koldo Miguel et al., « Home Camera-Based Fall Detection System for the Elderly », in : *Sensors* 17 (déc. 2017), p. 2864, doi : 10.3390/s17122864.

[34] Jeffrey Kutchka et al., « Automatic Assessment of Environmental Hazards for Fall Prevention Using Smart-Cameras », in : *IEEE First International Conference on Connected Health : Applications, Systems and Engineering Technologies (CHASE)*, juin 2016, p. 24-29, doi : 10.1109/CHASE.2016.53.

[35] Kaibo Fan et al., « Fall detection via human posture representation and support vector machine », in : *International Journal of Distributed Sensor Networks* 13 (mai 2017), doi : 10.1177/1550147717707418.

[36] Tzung-Han Lin, Chi-Yun Yang et Shih Wen-Pin, « Fall Prevention Shoes Using Camera-Based Line-Laser Obstacle Detection System », in : *Journal of Healthcare Engineering* 2017 (mai 2017), p. 1-11, doi : 10.1155/2017/8264071.

[37] Feng Zhao et al., « Real-Time Detection of Fall From Bed Using a Single Depth Camera », in : *IEEE Transactions on Automation Science and Engineering* PP (août 2018), p. 1-15, DOI : 10.1109/TASE.2018.2861382.

[38] Min Li et al., « Pre-impact fall detection based on a modified zero moment point criterion using data from Kinect sensors », in : *IEEE Sensors Journal* 18 (mai 2018), p. 5522-5531, DOI : 10.1109/JSEN.2018.2833451.

[39] Xiangbo Kong, Lin Meng et Hiroyuki Tomiyama, « Fall detection for elderly persons using a depth camera », in : *International Conference on Advanced Mechatronic Systems (ICAMechS)*, déc. 2017, p. 269-273, DOI : 10.1109/ICAMechS.2017.8316483.

[40] Tao Xu et Yun Zhou, « Fall prediction based on biomechanics equilibrium using Kinect », in : *International Journal of Distributed Sensor Networks* 13 (avr. 2017), p. 155014771770325, DOI : 10.1177/1550147717703257.

[41] Erdem Akagündüz et al., « Silhouette Orientation Volumes for Efficient Fall Detection in Depth Videos », in : *IEEE Journal of Biomedical and Health Informatics* 21 (mai 2016), p. 1-1, DOI : 10.1109/JBHI.2016.2570300.

[42] Amandine Dubois et François Charpillet, « A Gait Analysis Method Based on a Depth Camera for Fall Prevention », in : *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, t. 2014, août 2014, p. 4515-8, DOI : 10.1109/EMBC.2014.6944627.

[43] Kyu-Seob Song, Young-Hoon Nho et Dong-Soo Kwon, « Histogram based fall prediction of patients using a thermal imagery camera », in : *14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, juin 2017, p. 161-164, DOI : 10.1109/URAI.2017.7992700.

[44] Kyu-Seob Song, Young-Hoon Nho et Dong-Soo Kwon, « Histogram based fall prediction of patients using a thermal imagery camera », in : juin 2017, p. 161-164, DOI : 10.1109/URAI.2017.7992700.

[45] J.A. Stevens et al., « The Costs of Fatal and Non-Fatal Falls among Older Adults », in : *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention* 12 (oct. 2006), p. 290-5, DOI : 10.1136/ip.2005.011015.

[46] Frédéric Bloch et al., « Can metabolic abnormalities after a fall predict short term mortality in elderly patients ? », in : *European journal of epidemiology* 24 (fév. 2009), p. 357-62, DOI : 10.1007/s10654-009-9342-y.

[47] Geoffroy CORMIER, « Analyse statique et dynamique de cartes de profondeurs. Application au suivi des personnes à risque sur le lieu de vie. », thèse de doct., Uiversité de Rennes I, 2015.

[48] Tong Jia et al., « Depth Measurement Based on Infrared Coded Structured Light », in : *Hongwai yu Jiguang Gongcheng Infrared and Laser Engineering* 44 (mai 2015), p. 1628-1632, DOI : 10.1155/2014/852621.

[49] Loren Arthur Schwarz et al., « Human skeleton tracking from depth data using geodesic distances and optical flow », in : *Image and Vision Computing* 30.*3* (2012), Best of Automatic Face and Gesture Recognition 2011, p. 217-226, ISSN : 0262-8856, DOI : https://doi.org/10.1016/j.imavis.2011.12.001, URL : http://www.sciencedirect.com/science/article/pii/S026288561100134X.

[50] Leandro Cruz, Djalma Lucio et Luiz Velho, « Kinect and RGBD Images : Challenges and Applications », in : *25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, août 2012, DOI : 10.1109/SIBGRAPI-T.2012.13.

[51] Salma Jarray, « Computer Vision Based Fall Detection Methods Using the Kinect Camera : A Survey », in : *International Journal of Computer Science and Information Technology* 10 (oct. 2018), p. 73-92, DOI : 10.5121/ijcsit.2018.10507.

[52] *Put slam*, http://lrm.put.poznan.pl/putslam/, Accessed : 2019-10-16.

[53] Bogdan Kwolek et Michal Kepski, « Human fall detection on embedded platform using depth maps and wireless accelerometer », in : (jan. 2014).

[54] Erik Stone et Marjorie Skubic, « Fall Detection in Homes of Older Adults Using the Microsoft Kinect », in : *IEEE journal of biomedical and health informatics* 19 (mar. 2014), DOI : 10.1109/JBHI.2014.2312180.

[55] *API Reference libfreenect2*, https://openkinect.github.io/libfreenect2/, Accessed : 2017-10-16.

[56] Xin Ma et al., « Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine », in : *IEEE Journal of Biomedical and Health Informatics* PP (nov. 2014), p. 1-9, DOI : 10.1109/JBHI.2014.2304357.

[57] Georgios Mastorakis et Dimitrios Makris, « Fall detection system using Kinects infrared sensor », in : *Journal of Real-Time Image Processing* 9 (déc. 2014), doi : 10.1007/s11554-012-0246-9.

[58] Fabián Riquelme et al., « eHomeSeniors Dataset : An Infrared Thermal Sensor Dataset for Automatic Fall Detection Research », in : *Sensors* 19 (oct. 2019), p. 4565, doi : 10.3390/s19204565.

[59] *TASDA - Protocole dévaluation expérimentale des performances de détecteurs de chute du marché*, https://www.tasda.fr/wp-content/uploads/2019/01/TASDA-EvaluationCapteursChute-Protocole.pdf,2013., Accessed : 2018-09-02.

[60] C. Harris et M. Stephens, « A Combined Corner and Edge Detector », in : *Procedings of the Alvey Vision Conference 1988*, Alvey Vision Club, 1988, doi : 10.5244/c.2.23, url : https://doi.org/10.5244%2Fc.2.23.

[61] Jagadeesh Basavaiah et Chandrashekar Patil, « Video Based Human Activity Detection, Recognition and Classification of actions using SVM », in : *Transactions on Machine Learning and Artificial Intelligence* 6 (déc. 2018), doi : 10.14738/tmlai.66.5287.

[62] Allah Bux Sargana, Plamen Angelov et Zulfiqar Habib, « Vision Based Human Activity Recognition : A Review », in : t. 513, jan. 2017, p. 341-371, isbn : 978-3-319-46561-6, doi : 10.1007/978-3-319-46562-3_23.

[63] K P Kumar et Bhavani R., « Human activity recognition in egocentric video using HOG, GiST and color features », in : *Multimedia Tools and Applications* 79 (mai 2018), doi : 10.1007/s11042-018-6034-1.

[64] Kong Yucai et al., « Person following based on Haar-like feature and HOG feature in indoor environment », in : *35th Chinese Control Conference (CCC)* (2016), p. 6345-6349.

[65] Anshuman Agarwal, Shivam Gupta et Dushyant Singh, « Review of optical flow technique for moving object detection », in : *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, déc. 2016, p. 409-413, doi : 10.1109/IC3I.2016.7917999.

[66] Berthold Horn et Brian Schunck, « Determining Optical Flow », in : *Artificial Intelligence* 17 (août 1981), p. 185-203, doi : 10.1016/0004-3702(81)90024-2.

[67] Bruce Lucas et Takeo Kanade, « An Iterative Image Registration Technique with an Application to Stereo Vision », in : *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, t. 81, avr. 1981.

[68] Jiang Gao et al., « Dining Activity Analysis Using a Hidden Markov Model », in : *Proceedings of the 17th International Conference on Pattern Recognition*, jan. 2004, p. 915-918, doi : 10.1109/ICPR.2004.1334408.

[69] Frédéric Jean, Robert Bergevin et Alexandra Albu, « Body tracking in human walk from monocular video sequences », in : *2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, mai 2005, p. 144-151, isbn : 0-7695-2319-6, doi : 10.1109/CRV.2005.24.

[70] T. Senst, V. Eiselein et T. Sikora, « Robust Local Optical Flow for Feature Tracking », in : *IEEE Transactions on Circuits and Systems for Video Technology* 22.*9* (2012), p. 1377-1387.

[71] Y. Hu, R. Song et Y. Li, « Efficient Coarse-to-Fine Patch Match for Large Displacement Optical Flow », in : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 5704-5712.

[72] Renjie Li et Songyu Yu, « Confidence based optical flow algorithm for high reliability », in : *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), p. 785-788.

[73] G. Jemilda et S. Baulkani, « Moving Object Detection and Tracking using Genetic Algorithm Enabled Extreme Learning Machine », in : *International Journal of Computers Communications and Control* 13 (avr. 2018), p. 162-174, doi : 10.15837/ijccc.2018.2.3064.

[74] Klaus Greff et al., « A comparison between background subtraction algorithms using a consumer depth camera », in : *VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications*, t. 1, fév. 2012.

[75] Cannons Kevin, *A Review of Visual Tracking*, 2008.

[76] C. R. Wren et al., « Pfinder : real-time tracking of the human body », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.*7* (1997), p. 780-785.

[77] Eric Hayman et Jan-Olof Eklundh, « Statistical Background Subtraction for a Mobile Observer », in : *Proceedings Ninth IEEE International Conference on Computer Vision*, t. 1, nov. 2003, 67-74 vol.1, isbn : 0-7695-1950-4, doi : 10.1109/ICCV.2003.1238315.

[78] Zoran Zivkovic et F. Van der Heijden, « Efficient adaptive density estimation per image pixel for the task of background subtraction », in : *Pattern Recognition Letters* 27 (mai 2006), p. 773-780, doi : 10.1016/j.patrec.2005.11.005.

[79] Md. Delowar Hossain et Eui-Nam Huh, « Performance Assessment of Background Subtraction Algorithm », in : *KIISE Transactions on Computing Practices*, t. 25, déc. 2019, p. 344-350, doi : 10.5626/KTCP.2019.25.7.344.

[80] Anh Nghiem et Francois Bremond, « Background subtraction in people detection framework for RGB-D cameras », in : *11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, août 2014, p. 241-246, doi : 10.1109/AVSS.2014.6918675.

[81] Dongdong Zeng et al., « Background Subtraction With Real-Time Semantic Segmentation », in : *IEEE Access* (fév. 2019), p. 1-1, doi : 10.1109/ACCESS.2019.2899348.

[82] Miao Yu, Syed Naqvi et Jonathon Chambers, « Fall detection in the elderly by head tracking », in : *IEEE Workshop on Statistical Signal Processing Proceedings*, oct. 2009, p. 357-360, doi : 10.1109/SSP.2009.5278566.

[83] Caroline Rougier et al., « Monocular 3D Head Tracking to Detect Falls of Elderly People », in : *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 1 (fév. 2006), p. 6384-7, doi : 10.1109/IEMBS.2006.260829.

[84] Bart Jansen, Frederik Temmermans et Rudi Deklerck, « 3D human pose recognition for home monitoring of elderly », in : *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2007 (fév. 2007), p. 4049-51, doi : 10.1109/IEMBS.2007.4353222.

[85] Sigurjón Guðmundsson et al., « TOF imaging in Smart room environments towards improved people tracking », in : (juin 2008), doi : 10.1109/CVPRW.2008.4563154.

[86] Michael Harville, Gwynnelle Gordon et John Woodfill, « Foreground segmentation using adaptive mixture models in color anddepth », in : *Detection and Recognition of Events in Video*, fév. 2001, p. 3-11, isbn : 0-7695-1293-3, doi : 10.1109/EVENT.2001.938860.

[87] Caroline Rougier et al., « Fall Detection from Human Shape and Motion History Using Video Surveillance », in : *1st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, t. 2, juin 2007, p. 875-880, doi : 10.1109/AINAW.2007.181.

[88] Anh Nghiem, Edouard Auvinet et Jean Meunier, « Head detection using Kinect camera and its application to fall detection », in : *11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012*, juil. 2012, p. 164-169, isbn : 978-1-4673-0381-1, doi : 10.1109/ISSPA.2012.6310538.

[89] Siyuan Chen et al., « Exploring depth information for head detection with depth images », in : *13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, août 2016, p. 228-234, doi : 10.1109/AVSS.2016.7738060.

[90] Diego Ballotta et al., « Head Detection with Depth Images in the Wild », in : (juil. 2017).

[91] Diego Ballotta et al., « Fully Convolutional Network for Head Detection with Depth Images », in : *24th International Conference on Pattern Recognition (ICPR)*, août 2018, p. 752-757, doi : 10.1109/ICPR.2018.8545332.

[92] V. Vaidehi et al., « Video based automatic fall detection in indoor environment », in : (juin 2011), doi : 10.1109/ICRTIT.2011.5972252.

[93] Samuele Gasparrini et al., « Proposal and Experimental Evaluation of Fall Detection Solution Based on Wearable and Depth Data Fusion », in : *Advances in Intelligent Systems and Computing* 399 (jan. 2016), p. 99-108, doi : 10.1007/978-3-319-25733-4_11.

[94] Zhenpeng Bian, Lap-Pui Chau et Nadia Thalmann, « Fall detection based on skeleton extraction », in : *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, déc. 2012, p. 91-94, doi : 10.1145/2407516.2407544.

[95] Min Li et al., « Pre-impact fall detection based on a modified zero moment point criterion using data from Kinect sensors », in : *IEEE Sensors Journal* 18 (mai 2018), p. 5522-5531, DOI : 10.1109/JSEN.2018.2833451.

[96] D. Bansal et al., « Elderly People Fall Detection System Using Skeleton Tracking and Recognition », in : *American Journal of Applied Sciences* 15 (sept. 2018), p. 423-431, DOI : 10.3844/ajassp.2018.423.431.

[97] Jia-Luen Chua, Yoong Chang et Wee Lim, « A simple vision-based fall detection technique for indoor video surveillance », in : *Signal, Image and Video Processing* 9 (mar. 2013), DOI : 10.1007/s11760-013-0493-7.

[98] Ahmed Abobakr, Mo Hossny et Saeid Nahavandi, « A Skeleton-Free Fall Detection System From Depth Images Using Random Decision Forest », in : *IEEE Systems Journal* PP (déc. 2017), p. 1-12, DOI : 10.1109/JSYST.2017.2780260.

[99] Amin Amini, Konstantinos Banitsas et John Cosmas, « A comparison between heuristic and machine learning techniques in fall detection using Kinect v2 », in : *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, mai 2016, p. 1-6, DOI : 10.1109/MeMeA.2016.7533763.

[100] Bart Jansen et Rudi Deklerck, « Context aware inactivity recognition for visual fall detection », in : *Pervasive Health Conference and Workshops*, jan. 2007, p. 1-4, DOI : 10.1109/PCTHEALTH.2006.361657.

[101] Yueng Delahoz et Miguel Labrador, « Survey on Fall Detection and Fall Prevention Using Wearable and External Sensors », in : *Sensors (Basel, Switzerland)* 14 (oct. 2014), p. 19806-19842, DOI : 10.3390/s141019806.

[102] Xinguo Yu, « Approaches and principles of fall detection for elderly and patient », in : *10th IEEE Intl. Conf. on e-Health Networking, Applications and Service, HEALTHCOM 2008*, août 2008, p. 42-47, ISBN : 978-1-4244-2280-7, DOI : 10.1109/HEALTH.2008.4600107.

[103] Amandine Dubois et François Charpillet, « Human Activities Recognition with RGB-Depth Camera using HMM », in : *EMBC - 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society - 2013*, Osaka, Japan, juil. 2013, DOI : 10.1109/EMBC.2013.6610588, URL : https://hal.inria.fr/hal-00914319.

[104] D. Triantafyllou et al., « A Real-time Fall Detection System for Maintenance Activities in Indoor Environments**This work has been partially supported by the European Commission through the project HORIZON 2020-INNOVATION ACTIONS (IA)-636302-SATISFACTORY. », in : *IFAC-PapersOnLine* 49.*28* (2016), 3rd IFAC Workshop on Advanced Maintenance Engineering, Services and Technology AMEST 2016, p. 286-290, ISSN : 2405-8963, DOI : https://doi.org/10.1016/j.ifacol.2016.11.049, URL : http://www.sciencedirect.com/science/article/pii/S2405896316324752.

[105] G Welch et G Bishop, « An introduction to the Kalman filter (Technical Report No. TR 95-041) », in : (jan. 1995).

[106] Z. Qiu et al., « An adaptive Kalman predictor applied to tracking vehicles in the traffic monitoring system », in : *IEEE Proceedings. Intelligent Vehicles Symposium*, juil. 2005, p. 230-235, ISBN : 0-7803-8961-1, DOI : 10.1109/IVS.2005.1505107.

[107] Li Liang-qun, Ji Hong-bing et Luo Jun-hui, « The iterated extended Kalman particle filter », in : *International Symposium on Communications and Information Technology*, t. 34, nov. 2005, p. 1213-1216, ISBN : 0-7803-9538-7, DOI : 10.1109/ISCIT.2005.1567087.

[108] Shovan Bhaumik et Paresh Date, *The Kalman filter and the extended Kalman filter*, Nonlinear Estimation, juil. 2019, p. 27-50, ISBN : 9781351012355, DOI : 10.1201/9781351012355-2.

[109] Przemysaw Pasek et Piotr Kaniewski, « Unscented Kalman filter application in personal navigation », in : *Radioelectronic Systems Conference 2019*, sous la dir. de Piotr Kaniewski et Jan Matuszewski, t. 11442, International Society for Optics et Photonics, SPIE, 2020, p. 446-453, DOI : 10.1117/12.2564984, URL : https://doi.org/10.1117/12.2564984.

[110] Paulo Salgado et Paulo Afonso, « Body Fall Detection with Kalman Filter and SVM », in : *Lecture Notes in Electrical Engineering* 321 (jan. 2015), p. 407-416, DOI : 10.1007/978-3-319-10380-8_39.

[111] Michael Stevens et al., « A Kalman filter to estimate altitude change during a fall », in : *International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, t. 2016, août 2016, p. 5889-5892, DOI : 10.1109/EMBC.2016.7592068.

[112] Long Jiao et al., « Anatomy of a multicamera video surveillance system », in : *Multimedia Syst.* 10 (août 2004), p. 144-163, DOI : 10.1007/s00530-004-0147-2.

[113] Tai-shan Lou et al., « Rank Kalman Filter-SLAM for Vehicle with Non-Gaussian Noise », in : *5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, déc. 2020, p. 12-15, DOI : 10.1109/ICARM49381.2020.9195299.

[114] Caroline Rougier, « Vidéosurveillance intelligente pour la détection de chutes chez les personnes âgées. », thèse de doct., Faculté des arts et des sciences, Université de Montréal, 2010.

[115] Michael Isard et Andrew Blake, « CONDENSATIONconditional density propagation for visual tracking », in : *International Journal of Computer Vision* 29 (août 1998), p. 5-28, DOI : 10.1023/A:1008078328650.

[116] Hammadi Nait-Charif et Stephen Mckenna, « Activity summarisation and fall detection in a supportive home environment », in : *International Conference on Pattern Recognition*, t. 4, sept. 2004, 323-326 Vol.4, ISBN : 0-7695-2128-2, DOI : 10.1109/ICPR.2004.1333768.

[117] Katja Nummiaro et al., « Color-Based Object Tracking in Multi-camera Environments », in : *Lecture Note in Computer Science*, t. 2781, sept. 2003, p. 591-599, ISBN : 978-3-540-40861-1, DOI : 10.1007/978-3-540-45243-0_75.

[118] A. Treptow, G. Cielniak et Tom Duckett, « Active people recognition using thermal and grey images on a mobile security robot », in : *RSJ International Conference on Intelligent Robots and Systems*, sept. 2005, p. 2103-2108, DOI : 10.1109/IROS.2005.1545530.

[119] Robert M. Haralick et Linda G. Shapiro, *Computer and Robot Vision*, 1st, USA : Addison-Wesley Longman Publishing Co., Inc., 1992, ISBN : 0201569434.

[120] Caroline Rougier et al., « Robust Video Surveillance for Fall Detection Based on Human Shape Deformation », in : *Circuits and Systems for Video Technology, IEEE Transactions on* 21 (juin 2011), p. 611-622, DOI : 10.1109/TCSVT.2011.2129370.

[121] Caroline Rougier et al., « 3D head tracking for fall detection using a single calibrated camera », in : *Image and Vision Computing* 31.*3* (2013), p. 246-254, ISSN : 0262-8856, DOI : https://doi.org/10.1016/j.imavis.2012.11.003, URL : http://www.sciencedirect.com/science/article/pii/S0262885612002107.

[122] Stan. Birchfield, « Elliptical head tracking using intensity gradients and color histograms », in : *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, 1998, p. 232-237.

[123] Yi Wu, Jongwoo Lim et Ming-Hsuan Yang, « Online Object Tracking : A Benchmark », in : *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, juin 2013, p. 2411-2418, doi : 10.1109/CVPR.2013.312.

[124] Md Islam et al., « Deep Learning Based Systems Developed for Fall Detection : A Review », in : *IEEE Access* 8 (sept. 2020), doi : 10.1109/ACCESS.2020.3021943.

[125] Waseem Rawat et Zenghui Wang, « Deep Convolutional Neural Networks for Image Classification : A Comprehensive Review », in : *Neural Computation* 29 (juin 2017), p. 1-98, doi : 10.1162/NECO_a_00990.

[126] Athanasios Voulodimos et al., « Deep Learning for Computer Vision : A Brief Review », in : *Computational Intelligence and Neuroscience* 2018 (fév. 2018), p. 1-13, doi : 10.1155/2018/7068349.

[127] D HUBEL et T WIESEL, « Receptive fields, binocular interaction and functional architectures in cats visual cortex », in : *The Journal of physiology* 160 (fév. 1962), p. 106-54, doi : 10.1113/jphysiol.1962.sp006837.

[128] Walter Pinaya et al., « Convolutional neural networks », in : *Machine Learning : Methods and Applications to Brain Disorders*, nov. 2019, p. 17, isbn : 9780128157398, doi : 10.1016/B978-0-12-815739-8.00010-9.

[129] Ravishankar Chityala et Sridevi Pudipeddi, « Convolutional Neural Network », in : *Image Processing and Acquisition using Python*, juin 2020, p. 265-273, isbn : 9780429243370, doi : 10.1201/9780429243370-12.

[130] S. Albawi, T. A. Mohammed et S. Al-Zawi, « Understanding of a convolutional neural network », in : *2017 International Conference on Engineering and Technology (ICET)*, 2017, p. 1-6, doi : 10.1109/ICEngTechnol.2017.8308186.

[131] Keiron O'Shea et Ryan Nash, « An Introduction to Convolutional Neural Networks », in : *ArXiv e-prints* (nov. 2015).

[132] Rikiya Yamashita et al., « Convolutional neural networks : an overview and application in radiology », in : *Insights into Imaging* 9 (juin 2018), doi : 10.1007/s13244-018-0639-9.

[133] Xiaopeng Zhu et al., « LFN : Based on the Convolutional Neural Network of Gait Recognition Method », in : *Journal of Physics : Conference Series* 1650 (oct. 2020), p. 032075, DOI : 10.1088/1742-6596/1650/3/032075.

[134] Kripesh Adhikari, Hamid Bouchachia et Hammadi Nait-Charif, « Activity recognition for indoor fall detection using convolutional neural network », in : *IAPR International Conference on Machine Vision Applications (MVA)*, mai 2017, p. 81-84, DOI : 10.23919/MVA.2017.7986795.

[135] Tiago Lima, Bruno Fernandes et Pablo Barros, « Human action recognition with 3D convolutional neural network », in : *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, nov. 2017, p. 1-6, DOI : 10.1109/LA-CCI.2017.8285700.

[136] Jun Chen et al., « An Improved Two-stream 3D Convolutional Neural Network for Human Action Recognition », in : *25th International Conference on Automation and Computing (ICAC)*, sept. 2019, p. 1-6, DOI : 10.23919/IConAC.2019.8894962.

[137] Jingmei Li et al., « An Improved Human Action Recognition Method Based on 3D Convolutional Neural Network », in : *Advanced Hybrid Information Processing*, mai 2019, p. 37-46, ISBN : 978-3-030-19085-9, DOI : 10.1007/978-3-030-19086-6_5.

[138] Karen Simonyan et Andrew Zisserman, « Two-Stream Convolutional Networks for Action Recognition in Videos », in : *Advances in Neural Information Processing Systems* 1 (juin 2014).

[139] Ajrun Jain et al., « Learning Human Pose Estimation Features with Convolutional Networks », in : *arXiv e-prints* (déc. 2013).

[140] Umer Rafi et al., « An Efficient Convolutional Network for Human Pose Estimation », in : *Proceedings of the British Machine Vision Conference (BMVC)*, jan. 2016, p. 109.1-109.11, DOI : 10.5244/C.30.109.

[141] Chenxu Luo, Xiao Chu et Alan Yuille, « OriNet : A Fully Convolutional Network for 3D Human Pose Estimation », in : *BMVC 2018 - Proceedings of the British Machine Vision Conference 2018*, nov. 2018.

[142] Adrián Núñez-Marcos, Gorka Azkune et Ignacio Arganda-Carreras, « Vision-Based Fall Detection with Convolutional Neural Networks », in : *Wireless Communications and Mobile Computing* 2017 (déc. 2017), p. 1-16, DOI : 10.1155/2017/9474806.

[143] Baopu Li et al., « Indoor human detection using RGB-D images », in : *IEEE International Conference on Information and Automation (ICIA)*, août 2016, p. 1354-1360, DOI : 10.1109/ICInfA.2016.7832030.

[144] Miao Yu, Liyun Gong et Stefanos Kollias, « Computer vision based fall detection by a convolutional neural network », in : *ICMI '17 : Proceedings of the 19th ACM International Conference on Multimodal Interaction*, nov. 2017, p. 416-420, DOI : 10.1145/3136755.3136802.

[145] Anahita Shojaei-Hashemi et al., « Video-based Human Fall Detection in Smart Homes Using Deep Learning », in : *IEEE International Symposium on Circuits and Systems (ISCAS)*, mai 2018, p. 1-5, DOI : 10.1109/ISCAS.2018.8351648.

[146] Markus Solbach et John Tsotsos, « Vision-Based Fallen Person Detection for the Elderly », in : *IEEE International Conference on Computer Vision Workshops (ICCVW)*, oct. 2017, p. 1433-1442, DOI : 10.1109/ICCVW.2017.170.

[147] Quanzeng You et al., « Image Captioning with Semantic Attention », in : *CVPR'16 Computer Vision and Pattern Recognition*, juin 2016, p. 4651-4659, DOI : 10.1109/CVPR.2016.503.

[148] Ross Girshick et al., « Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation », in : *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (nov. 2013), DOI : 10.1109/CVPR.2014.81.

[149] Kaiming He et al., « Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (juin 2014), DOI : 10.1109/TPAMI.2015.2389824.

[150] Ross Girshick, « Fast r-cnn », in : *IEEE International Conference on Computer Vision (ICCV)*, avr. 2015, DOI : 10.1109/ICCV.2015.169.

[151] Shaoqing Ren et al., « Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (juin 2015), DOI : 10.1109/TPAMI.2016.2577031.

[152] Jifeng Dai et al., « R-fcn : Object detection via region-based fully convolutional networks », in : *Computer Vision and Pattern Recognition*, mai 2016.

[153] Tsung-Yi Lin et al., « Feature Pyramid Networks for Object Detection », in : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, déc. 2016.

[154] Kaiming He et al., « Mask R-CNN », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (juin 2018), p. 1-1, DOI : 10.1109/TPAMI.2018.2844175.

[155] Dumitru Erhan et al., « Scalable Object Detection Using Deep Neural Networks », in : (déc. 2013), DOI : 10.1109/CVPR.2014.276.

[156] Jiaxing Mao, Hao Dou et Jinwen Tian, « G-CNN type recognition of typical aircraft based on target characteristics », in : *Eleventh International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2019)*, fév. 2020, p. 19, DOI : 10.1117/12.2537974.

[157] Joseph Redmon et al., « You Only Look Once : Unified, Real-Time Object Detection », in : 2016, arXiv : 1506.02640 [cs.CV].

[158] Wei Liu et al., « SSD : Single Shot MultiBox Detector », in : Springer International Publishing, 2016, p. 21-37, ISBN : 9783319464480, DOI : 10.1007/978-3-319-46448-0_2, URL : http://dx.doi.org/10.1007/978-3-319-46448-0_2.

[159] Joseph Redmon et Ali Farhadi, « YOLO9000 : Better, Faster, Stronger », in : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, juil. 2017, p. 6517-6525, DOI : 10.1109/CVPR.2017.690.

[160] Jasper Uijlings et al., « Selective Search for Object Recognition », in : *International Journal of Computer Vision* 104 (sept. 2013), p. 154-171, DOI : 10.1007/s11263-013-0620-5.

[161] Alexandru Telea, « An Image Inpainting Technique Based on the Fast Marching Method », in : *Journal of Graphics Tools* 9 (jan. 2004), DOI : 10.1080/10867651.2004.10487596.

[162] PoLin Lai, Dong Tian et Patrick Lopez, « Depth map processing with iterative joint multilateral filtering », in : *28th Picture Coding Symposium, PCS 2010*, jan. 2011, p. 9-12, DOI : 10.1109/PCS.2010.5702589.

[163] Junyi Liu, Xiaojin Gong et Jilin Liu, « Guided inpainting and filtering for Kinect depth maps », in : *International Conference on Pattern Recognition*, jan. 2012, p. 2055-2058, ISBN : 978-1-4673-2216-4.

[164] Lai Po et al., « A new multidirectional extrapolation hole-filling method for Depth-Image-Based Rendering », in : *Proceedings - International Conference on Image Processing, ICIP*, sept. 2011, p. 2589-2592, DOI : 10.1109/ICIP.2011.6116194.

[165] Valeria Garro, Pietro Zanuttigh et Guido Cortelazzo, « A Novel Interpolation Scheme for Range Data with Side Information », in : *2010 Conference on Visual Media Production* 0 (nov. 2009), p. 52-60, doi : 10.1109/CVMP.2009.26.

[166] Sergey Matyunin et al., « Temporal filtering for depth maps generated by Kinect depth camera », in : *3DTV Conference : The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)* (mai 2011), p. 1-4, doi : 10.1109/3DTV.2011.5877202.

[167] Massimo Camplani et Luis Salgado, « Efficient Spatio-Temporal Hole Filling Strategy for Kinect Depth Maps », in : *Proc SPIE* 8290 (fév. 2012), p. 13-, doi : 10.1117/12.911909.

[168] Na-Eun Yang, Yong-Gon Kim et Rae-Hong Park, « Depth hole filling using the depth distribution of neighboring regions of depth holes in the Kinect sensor », in : *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, août 2012, p. 658-661, isbn : 978-1-4673-2192-1, doi : 10.1109/ICSPCC.2012.6335696.

[169] Ismaeel Daribo, Christophe Tillier et Beatrice Pesquet, « Distance Dependent Depth Filtering in 3D Warping for 3DTV », in : *IEEE 9th Workshop on Multimedia Signal Processing*, nov. 2007, p. 312-315, isbn : 978-1-4244-1274-7, doi : 10.1109/MMSP.2007.4412880.

[170] Xuyuan Xu et al., « Depth map misalignment correction and dilation for DIBR view synthesis », in : *Signal Processing Image Communication* 28 (oct. 2013), doi : 10.1016/j.image.2013.04.003.

[171] Y. Berdnikov et Dmitriy Vatolin, « Real-time Depth Map Occlusion Filling and Scene Background Restoration for Projected-Pattern-based Depth Camera », in : (sept. 2011), p. 1-4.

[172] Shahram Izadi et al., « KinectFusion : Real-time 3D reconstruction and interaction using a moving depth camera », in : *UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, oct. 2011, p. 559-568, doi : 10.1145/2047196.2047270.

[173] Christian Richardt et al., « Coherent Spatiotemporal Filtering, Upsampling and Rendering of RGBZ Videos », in : *Computer Graphics Forum* 31 (mai 2012), p. 247-256, doi : 10.1111/j.1467-8659.2012.03003.x.

[174] Abm Islam et al., « Robust Enhancement of Depth Images from Depth Sensors », in : *Computers AND Graphics* 68 (août 2017), DOI : 10.1016/j.cag.2017.08.003.

[175] Amir Atapour-Abarghouei et Toby P. Breckon, « A comparative review of plausible hole filling strategies in the context of scene depth image completion », in : *Computers AND Graphics* 72 (2018), p. 39-58, ISSN : 0097-8493, DOI : https://doi.org/10.1016/j.cag.2018.02.001, URL : http://www.sciencedirect.com/science/article/pii/S0097849318300219.

[176] Karen Simonyan et Andrew Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », in : *arXiv 1409.1556* (sept. 2014).

[177] Kaiming He et al., « Deep Residual Learning for Image Recognition », in : juin 2016, p. 770-778, DOI : 10.1109/CVPR.2016.90.

[178] Debojit Biswas et al., « An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD », in : *Physics and Chemistry of the Earth, Parts A/B/C* 110 (2019), Sensing and Sensor Systems for Urban Environmental Studies, p. 176-184, ISSN : 1474-7065, DOI : https://doi.org/10.1016/j.pce.2018.12.001, URL : http://www.sciencedirect.com/science/article/pii/S1474706518302389.

[179] Pierre Sermanet et al., « OverFeat : Integrated Recognition, Localization and Detection using Convolutional Networks », in : *International Conference on Learning Representations (ICLR) (Banff)* (déc. 2013).

[180] Shuren Zhou et Jia Qiu, « Improved SSD for Object Detection », in : *Artificial Intelligence and Security*, sept. 2020, p. 313-322, ISBN : 978-981-15-8082-6, DOI : 10.1007/978-981-15-8083-3_28.

[181] Ross Girshick, « Fast r-cnn », in : *IEEE International Conference on Computer Vision (ICCV)*, avr. 2015, DOI : 10.1109/ICCV.2015.169.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Suivi de l'activité d'une personne à partir de capteurs multi-modalités préservant l'anonymat dans un cadre de détection et prévention des chutes chez les personnes âgées

Mot clés : Suivi de la personne ; Fusion ; Reconnaissance de l'activité ; Deep learning ; Thermique ; Profondeur

Résumé : La sécurité des personnes âgées est un enjeu majeur de santé publique. Les travaux de cette Thèse ont porté plus particulièrement un système de détection et de prévention des chutes pouvant être utilisé en Ehpad ou au domicile des personnes autonomes. Pour cela, nous avons utilisé un dispositif à bas coût composé d'une caméra de profondeur et d'une caméra thermique. Ce dispositif fonctionne d'une façon autonome, en temps réel, de jour comme de nuit et permet de préserver l'anonymat de la personne. L'objectif de la thèse était de développer des algorithmes de suivi de la personne et d'identification de son activité afin d'extraire, à court terme, des paramètres pour la détection de la chute et, à long terme, des indices de la fragilité de la personne. Pour cette Thèse, nous avons exploré deux stratégies permettant ce suivi : 1) Une méthode classique de suivi de la personne basée sur du filtrage particulaire. Dans cette approche, les informations acquises par les deux caméras ont été fusionnées afin de suivre la tête de la personne tout au long de la séquence d'images. 2) Une méthode de reconnaissance de l'activité de la personne à partir dun apprentissage profond sur les séquences d'images thermiques et de profondeurs. Nous avons choisi de détecter quatre postures (assis, debout, allongé sur le lit et allongé par terre) à partir de nos propres bases de données. Nous pensons qu'a terme, l'analyse de la séquence temporelle des postures extraites par nos algorithmes permettra de détecter les signes de fragilité de la personne âgés.

Title: Tracking person activity using thermal and depth sensors in a fall detection and prevention context

Keywords: Person tracking; Fusion, Activity recognition; Deep learning; Thermal; Depth

Abstract: The safety of the elderly people is a major public health issue. The work of this Thesis concerned more particularly a fall detection and prevention system that can be used in retirement homes or in the home of autonomous people. For this, we used a low-cost device composed of a depth camera and a thermal camera. This device works autonomously, in real time, day and night and allows to preserve the privacy of the person. The objective of the thesis was to develop algorithms for tracking the person and identifying his activity in order to extract, in the short term, parameters for the detection of the fall and, in the long term, indices of the person's fragility. For this Thesis, we explored two strategies allowing this monitoring: 1) A people tracking method based on particle filtering. In this approach, the information acquired by the two cameras was merged in order to follow the per-

son's head throughout the image sequence. 2) A method of recognizing the person's activity based on deep learning applied on thermal and depth image sequences. We chose to detect four postures (sitting, standing, lying down and fall) from an own acquired dataset. We believe that in the long term, the analysis of the temporal sequence of the postures extracted by our algorithms will allow to detect signs of frailty of the elderly person.