REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU

FACULTE GENIE ELECTRIQUE ET INFORMATIQUE
DEPARTEMENT AUTOMATIQUE

Thèse

pour l'obtention du diplôme de

Doctorat 3ème Cycle LMD en Automatique

Thème

Identification automatique de silhouettes humaines dans des séquences d'images infrarouges

présentée par

Merzouk YOUNSI

Soutenue publiquement le 20 / 12 / 2023 devant le jury :

Rachid AHMED-OUAMAR	Professeur à l'UMMTO	Président
Moussa DIAF	Professeur à l'UMMTO	Rapporteur
Sarah BENZIANE	Professeur à l'USTO	Examinatrice
Slimane LARABI	M.C.A à l'USTHB	Examinateur
Sadia ALKAMA	M.C.A à l'UMMTO	Examinatrice

Année universitaire : 2023-2024

Avant-propos

La présente thèse a été réalisée au sein du Laboratoire Vision Artificielle et Automatique des Systèmes (LVAAS) de la faculté Génie électrique et Informatique, Département Automatique, de l'Université Mouloud Mammeri de Tizi-Ouzou (UMMTO).

Je tiens à remercier, en premier lieu, Monsieur **Moussa DIAF**, Professeur à l'université Mouloud Mammeri de Tizi-Ouzou, Directeur de ma thèse pour avoir dirigé ce travail de recherche tout au long de ces années de thèse. Qu'il me soit permis de lui exprimer toute ma reconnaissance et ma gratitude pour ses précieux conseils, sa disponibilité, sa patience et sa gentillesse et pour la bienveillante attention dont il m'a entouré.

Je tiens à remercier également Monsieur **Rachid AHMED-OUAMAR**, Professeur à l'Université Mouloud Mammeri de Tizi-Ouzou (UMMTO) de m'avoir fait l'honneur d'accepter de présider le jury de ma soutenance de thèse.

Mes remerciements vont également à Monsieur **Slimane LARABI**, Professeur à l'Université des Sciences et de la Technologie Houari-Boumédiène (USTHB), qui a accepté de faire partie du jury et porter un jugement à cette thèse.

Je souhaite également remercier Madame **Sarah BENZIANE**, Maitre de Conférences A à l'Université des Sciences et de la Technologie Oran Mohamed-Boudiaf (USTO MB), pour avoir accepté de faire partie de mon jury de thèse et accepté d'évaluer ce travail.

J'adresse aussi mes remerciements à Madame **Sadia ALKAMA Ép. HAMMOUCHE**, Maitre de Conférences A à l'Université Mouloud Mammeri de Tizi-Ouzou (UMMTO), pour avoir accepté de considérer et d'évaluer notre travail de recherche.

Mes remerciements vont également à tous mes collègues du laboratoire LVAAS, notamment Samir YESLI, Salim IRATNI et Abderezak SALMI avec lesquels j'ai passé toutes ces années en leur compagnie. L'ambiance dans laquelle nous avons travaillé et l'échange de discussion avec eux m'ont été d'une grande aide morale et scientifique.

A

- ✓ *la mémoire de mes défunts grands-parents qui n'ont pas pu voir ce travail,*
- ✓ *mes très chers parents pour leur amour parental et leurs encouragements,*
- ✓ *mon frère, sa femme et leurs enfants,*
- ✓ *mes sœurs, pour toute l'aide précieuse qu'elles m'ont-apportée tout au long de mes études,*
- ✓ *tous mes amis et collègues.*

Merzouk

Résumé

La détection, le suivi, et la reconnaissance de postures de personnes en mouvement est un sujet de recherche très actif dans le domaine de vision par ordinateur en raison de ses applications dans divers domaines tels que la vidéo surveillance, l'interaction homme-robot, la récupération vidéo, les soins de santé, les véhicules intelligents, la réalité virtuelle et la réalité augmentée. Cependant, le développement d'un système efficace pour la détection et le suivi de personnes reste une tâche difficile à traiter, notamment lorsque nous avons à faire face à des environnements extérieurs à faible luminosité tels que la nuit. D'autres facteurs rendant cette tâche plus ardue incluent : la présence dans la scène d'objets non-humains, les encombrements d'arrière-plan, les occultations, les changements d'apparence, les changements de postures, le bruit et la contrainte du temps réel. Afin de surmonter certaines de ces difficultés, dans ce travail de thèse, nous proposons un nouveau système de vidéo surveillance capable de détecter, suivre, et reconnaître efficacement la posture de personnes en mouvement à partir de séquences d'images acquises par une caméra infrarouge dans des environnements extérieurs de nuit. Ce système comprend les étapes suivantes. Après l'extraction des objets en mouvement en utilisant la méthode de soustraction d'arrière-plan, nous proposons deux approches différentes pour distinguer un être humain de toute autre forme d'objet en mouvement. La première approche est basée sur le calcul d'une fonction de similarité combinée qui utilise des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement. La seconde approche est basée sur la détection conjointe de deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre majuscule de l'alphabet grec Omega Ω), et les deux jambes. Une fois qu'un être humain est détecté, afin de le suivre efficacement et de manière robuste en cas de présence des situations difficiles citées précédemment, nous proposons une méthode qui utilise un filtre à particules et une combinaison adaptative d'informations provenant de plusieurs types de caractéristiques, à savoir l'intensité, la texture, la vélocité de mouvement, et la distance spatiale. Pour augmenter davantage la robustesse de notre méthode de suivi, nous introduisons aussi une stratégie automatique de détection et de traitement des occultations basée sur des règles heuristiques simples et l'histogramme de projection verticale en niveaux de gris. En parallèle avec l'algorithme de suivi, et pour décrire efficacement la posture de l'être humain détecté, nous extrayons de sa silhouette trois caractéristiques différentes, à savoir des caractéristiques basées région, basées contour et géométriques. Les performances de notre système proposé sont évaluées sur plusieurs séquences d'images infrarouges capturées dans des environnements réels de nuit, et les résultats expérimentaux obtenus ont démontré une bonne faisabilité et efficacité de notre système pour la détection automatique des êtres humains en mouvement et l'analyse de leurs comportements au cours du temps.

Mots clés

Images infrarouges, détection de personnes, suivi de cibles, reconnaissance de posture humaine, classification, analyse du comportement humain.

Abstract

Detection, tracking, and posture recognition of moving humans are very active research topics in the field of computer vision due to their applications in various domains such as video surveillance, human-robot interaction, video retrieval, healthcare, intelligent vehicles, virtual reality, and augmented reality. However, developing an effective system for human detection and tracking is still a challenging task to achieve, especially when we are dealing with low-visibility outdoor environments such as at night. Other factors that make this task more challenging include the presence of non-human objects in the scene, background clutter, occlusions, appearance changes, posture changes, noise, and the real-time constraint. To address some of these challenges, in this thesis, we propose a new video surveillance system capable of efficiently detecting, tracking, and recognizing the posture of moving humans from image sequences acquired by an infrared camera in outdoor night environments. This system includes the following steps. After extracting moving objects using the background subtraction method, we propose two different approaches to distinguish a human from any other form of moving objects. The first approach is based on the computation of a combined similarity function that uses shape, appearance, spatial and temporal information of the moving objects. The second approach is based on the joint detection of two parts that characterize the human body, namely the head-shoulder part (like-Omega Ω shape), and the two legs. Once a human is detected, to efficiently and robustly track it in the presence of the previously mentioned challenging situations, we propose a method that uses a particle filter and an adaptive combination of information from several types of features, namely intensity, texture, motion velocity, and spatial distance. To further increase the robustness of our tracking method, we also introduce an automatic strategy for occlusion detection and handling based on simple heuristic rules and the grayscale vertical projection histogram. In parallel with the tracking algorithm, and to effectively describe the posture of the detected human, we extract three different features from its silhouette, namely region-based features, contour-based features and geometric features. The performance of our proposed system is evaluated on several infrared image sequences captured in real-night environments, and the experimental results obtained have demonstrated the good feasibility and efficiency of our system for the automatic detection of moving humans and analysis of their behaviors over time.

Key words

Infrared images, human detection, target tracking, human posture recognition, classification, human behavior analysis.

Table des matières

Chapitre 3
Suivi de personnes

Chapitre 4
Reconnaissance de posture humaine

Chapitre 5
Résultats expérimentaux

Liste des figures

Liste de tableaux

Liste des publications et communications

Publications dans des revues à comité de lecture :

1. Younsi, M., Diaf, M., Siarry, P., 2020. Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera. Expert Systems with Applications, 146, 113171.

2. Younsi, M., Diaf, M., Siarry, P., 2023. Comparative study of orthogonal moments for human postures recognition. Engineering Applications of Artificial Intelligence, 120, 105855.

3. Younsi, M., Yesli, S., Diaf, M., 2023. Depth-based human action recognition using histogram of templates. Multimedia Tools and Applications, 1-35.

Communications lors de congrès internationaux :

1. Iratni, S., Younsi, M., Diaf, M., 2021. Human Detection Method based on Combined Face and Upper-Body Part Detectors, in: 4th international Conference on Artificial Vision (CVA' 2021).

Introduction Générale

1. Contexte du travail

Avec l'augmentation immense du nombre d'activités antisociales et de différents accidents dans la vie quotidienne, le recours à de systèmes de vidéosurveillance devient de plus en plus nécessaire et tout particulièrement pour sécuriser les habitations et des zones sensibles, telles que les aéroports, les banques, les bâtiments industriels, les frontières, les bâtiments militaires, gouvernementaux, etc. Cependant, la plupart des systèmes de vidéosurveillance existants actuellement nécessitent la présence d'un ou plusieurs opérateurs humains pour observer et analyser en permanence, en temps réel, les énormes quantités d'information en provenance des caméras, et ce, en vue de détecter une situation suspecte qui pourrait être à l'origine de dégâts portant atteinte à des personnes et à leurs biens. Dans le cas d'une sécurité accrue avec un nombre élevé d'écrans de surveillance, l'emploi de plusieurs opérateurs est important compte tenu de la fatigue et de la lassitude visuelle et psychique de l'être humain lorsque ce type de tâches se prolonge dans le temps. Comme conséquence, dans certains environnements difficiles et dangereux (plateformes pétrolières, sites miniers et métallurgiques, installations gazières, centrales nucléaires, etc.), l'utilisation de la vidéosurveillance dans ce cas devient très onéreuse et parfois réfutée. Ainsi, afin de remédier à ces inconvénients, le développement de systèmes de vidéosurveillance intelligents complètement automatiques, sans l'intervention d'un opérateur humain, est d'une nécessité majeure. C'est ainsi que, durant ces dernières années, une quantité considérable de recherches et de publications ont été consacrées au développement de ces algorithmes intelligents.

Pour répondre donc à ces besoins, ces systèmes de vidéosurveillance intelligents sont conçus sur la base d'algorithmes fiables, efficaces et rapides. La figure 0.1 montre un schéma général de fonctionnement d'un tel système dont les principales composantes comprennent les opérations d'extraction des objets en mouvement

(régions d'intérêt), de détection et de suivi de personnes, de reconnaissance de postures et d'analyse du comportement.

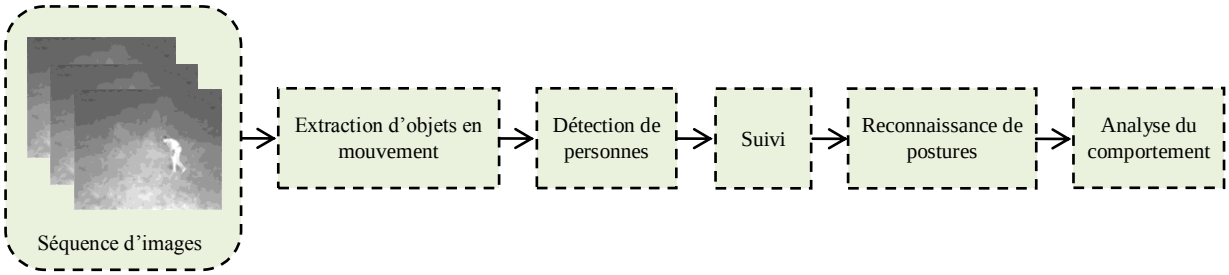


Figure 0.1 : Schéma général d'un système typique de vidéo surveillance intelligent.

L'extraction des objets en mouvement (régions d'intérêt) est l'étape fondamentale pour tout processus d'analyse approfondie des séquences d'images. Elle a pour objectif, la segmentation d'objets en mouvement à partir de l'arrière-plan statique de la scène observée. Cette étape, non seulement, elle restreint la zone de recherche des régions candidates pour la détection de personnes, mais elle permet aussi de réduire considérablement le temps de traitement. Pour cette opération, les techniques les plus couramment utilisées sont la soustraction d'arrière-plan, la différence temporelle et le flot optique. En raison des conditions dynamiques de l'environnement telles que les changements d'éclairage, les ombres et les branches d'arbre agitées par le vent, l'extraction des objets en mouvement reste un problème difficile et important auquel nous avons à faire face afin d'avoir un système de vidéo surveillance très robuste.

La deuxième étape dans l'analyse des séquences d'images acquises par des caméras de surveillance est la détection de personnes. Il s'agit de détecter la présence de personnes dans la scène observée et de déterminer leur emplacement tout en produisant le moins de fausses détections possibles. Actuellement, il existe deux types d'approche pour la détection de personnes (Nguyen et al., 2010): les approches basées sur la correspondance de modèles (*template matching-based approaches*) et les approches basées sur l'apprentissage (*learning-based approaches*). Dans les approches basées sur la correspondance de modèles, les humains sont décrits explicitement par des modèles représentant le corps dans son ensemble ou certains de ses membres (tels que la tête, le visage, les bras, etc.). La

tâche de détection de personnes consiste ainsi à trouver le meilleur appariement (best matching) entre les modèles et une région candidate de l'image d'entrée. Les modèles peuvent être représentés par des images d'intensité, ou de couleur, lorsque l'apparence des personnes est prise en compte ou simplement par des contours binaires lorsque des informations sur la forme sont utilisées. Dans les approches basées sur l'apprentissage, des caractéristiques d'apparence ou de forme sont tout d'abord extraites de l'image considérée, puis des algorithmes d'apprentissage supervisés, tels que les Machines à Vecteurs de Support (SVM), les réseaux de neurones artificiels et la classification bayésienne naïve (Naive Bayes classifier) sont utilisés pour créer un modèle de classification. Le problème de détection de personnes est alors souvent formulé sous la forme d'une classification binaire (humain/non humain). En raison des grandes variations dans la forme du corps d'une personne à une autre, des grands changements de posture chez chaque individu et à la trajectoire arbitraire des mouvements humains, la tâche de détection de personnes reste un problème difficile à traiter.

L'étape suivante dans le processus de vidéo surveillance intelligente est le suivi, qui peut être défini comme la mise en correspondance temporelle entre les personnes détectées dans des trames consécutives. Cette procédure permet une identification temporelle des régions segmentées et génère des informations cohérentes sur les personnes en mouvement dans la zone surveillée à savoir la trajectoire, la vitesse et la direction de mouvement. Le résultat de l'étape de suivi est généralement utilisé pour améliorer la qualité de segmentation de mouvement, la détection de personnes et l'analyse de comportements de haut niveau.

La reconnaissance de posture humaine joue un rôle très important dans le processus global de vidéo surveillance intelligente, car elle fournit des informations très importantes et utiles pour l'étape d'analyse du comportement. Son objectif est de reconnaître les différents types de postures qu'un être humain peut adopter, tels que "Debout" (Standing), "Assis" (Sitting), "Penché" (Bending), "Allongé" (Lying), etc. Deux types de techniques différentes sont couramment envisagés pour la reconnaissance de posture humaine (Boulay et al., 2006). Les premières sont des techniques dites intrusives et les deuxièmes des techniques dites non intrusives. Les techniques intrusives suivent généralement des marqueurs placés sur des points significatifs du corps (tels que les articulations, la tête, les mains, etc.) et ce, afin de reconnaître la posture d'une personne, alors que les techniques non

intrusives observent une personne avec une ou plusieurs caméras puis utilisent des algorithmes sophistiqués de traitement d'images et de vision par ordinateur pour effectuer la tâche de reconnaissance de posture. Cependant, la grande variété de postures dues au haut degré de liberté du corps humain ainsi que les apparences différentes que les personnes peuvent avoir dans une image, selon par exemple les vêtements ou les différents points de vue de la caméra, sont les principaux défis à relever lors du développement d'une méthode de reconnaissance de posture.

La dernière étape des systèmes de vidéo surveillance intelligents consiste à reconnaître le comportement des personnes suivies, et à créer une description de haut niveau de leurs actions. Cette tâche peut être considérée comme un problème de classification des signaux de l'activité temporelle des personnes suivies en fonction des signaux de référence pré-étiquetés représentant des actions humaines typiques (Wang et al., 2003).

Les résultats des différentes étapes décrites ci-dessus peuvent être utilisés pour fournir à l'opérateur humain des données de haut niveau afin de l'aider à prendre la décision précise et dans un temps plus court. Ces résultats peuvent être aussi utilisés pour l'indexation hors-ligne (sans contrainte de temps réel) et la recherche de vidéos par le contenu. Les progrès réalisés dans le développement de chacune des tâches ont apporté de nouvelles découvertes dans de nombreuses applications liées à la vidéo surveillance. Parmi les applications, on peut citer la sécurité publique, privée et militaire, l'extraction intelligente de données à partir des flux vidéo, etc. En effet, en sécurité publique, privée et militaire, il s'agit de surveiller les sites sensibles pour alerter sur des intrusions et contrôles d'accès (banques, aéroports, gares, stations de métro, musées, espaces commerciaux, parcs de stationnement, etc.) pour particulièrement détecter des actions dangereuses, des crimes et des accidents qui peuvent y survenir, y compris sur les autoroutes et les chemins de fer. Cette surveillance est appliquée aussi pour détecter le non-respect de l'arrêt au feu rouge, sécuriser des voies de circulation réservées aux véhicules (autoroutes, ponts, tunnels, etc.) ainsi que les zones strictement interdites aux piétons, telles que les pelouses et les espaces verts. Dans cette même catégorie, on peut inclure aussi la surveillance des forêts et des aires protégées afin de détecter des activités illégales, comme le braconnage, la chasse en période de fermeture, la coupe de bois, etc. Cette surveillance peut s'effectuer en temps réel et de manière

non intrusive comme dans le cas de suivi des activités des personnes âgées ou atteints de certaines maladies comme celles d'Alzheimer.
Le cas d'extraction intelligente de données des flux vidéo s'applique lorsqu'il s'agit de mesurer le flux de circulation et de la congestion des piétons, l'analyse du comportement des consommateurs dans les centres commerciaux et les parcs d'attraction, l'étude de la performance des sportifs et l'enregistrement des tâches de maintenance de routine dans les installations nucléaires et industrielles.
Dans le cas de la surveillance territoriale, il s'agit de la protection des frontières, de contrôler les flux migratoires et de sécuriser les périmètres et les accès autour des sites militaires.

2. Nos motivations

Comme mentionné précédemment, les tâches de détection de personnes, de suivi, et de reconnaissance automatique de posture humaine jouent un rôle très important dans un système de vidéo surveillance intelligente. Ainsi, au cours de ces dernières années, de nombreux systèmes dans ce domaine utilisant des images provenant de différents types de caméras ont été proposés. Les caméras standards sensibles à la lumière visible sont les plus largement utilisées dans les applications pratiques en raison de leur faible coût et leur capacité à fournir des images riches en informations. Parmi les nombreux systèmes basés sur des images issues de ce type de caméra, nous pouvons en citer ceux de (Wren et al., 1997; Collins et al., 2000; Haritaoglu et al., 2000; Wu and Nevatia, 2007; Benezeth et al., 2010; García-Martín and Martínez, 2012; Choi et al., 2015; Haq et al., 2020; Liu, 2021). Wu and Nevatia, (2007), par exemple, ont abordé le problème de détection de personnes en utilisant un ensemble de détecteurs de parties du corps appris par un renforcement (boosting) de plusieurs classifieurs faibles basés sur des caractéristiques extraites du contour, appelées "edgelets". Le suivi des personnes est réalisé par une méthode d'association de données combinée avec l'algorithme de mean-shift. Benezeth et al., (2010) ont effectué le suivi en combinant la méthode d'analyse en composantes connexes avec un suivi de points d'intérêt. La détection de personnes est réalisée avec plusieurs cascades de classifieurs boostés en utilisant l'algorithme AdaBoost et les caractéristiques peuso-Haar (Haar-like features). Choi et al., (2015) ont utilisé la modélisation et la soustraction d'arrière-plan pour extraire les régions d'intérêt, puis ils ont proposé une méthode à base de filtrage à particules pour le suivi de

personnes. L'algorithme glouton (greedy algorithm) a été utilisé pour faire correspondre les résultats détectés aux trackers actuels. Pour la tâche de reconnaissance de posture humaine à partir des caméras sensibles à la lumière visible, plusieurs approches basées sur différents types de caractéristiques ont également été proposées. Parmi les caractéristiques les plus couramment utilisées, on peut citer les histogrammes de projection horizontale et verticale (I. Haritaoglu et al., 2000; Goldmann et al., 2004; Cucchiara et al., 2005; Boulay et al., 2006; M. Yu et al., 2012), les moments de Hu (Boulay et al., 2006; Feng and Lin, 2010), le squelette (Collins et al., 2000; Chen et al., 2006; Boulay et al., 2006), et les opérations de transformation (Zerrouki and Houacine, 2014; Kang and Lee, 2016). L'un des principaux avantages des systèmes basés sur des images capturées par des caméras sensibles à la lumière visible est leur capacité à détecter et à suivre des personnes sur de très longues distances. De plus, comme la plupart de ces systèmes utilisent l'information de couleur, cela leur permet d'obtenir une description plus représentative et plus informative des cibles à suivre. Toutefois, ces systèmes sont souvent moins robustes face à des facteurs tels que les ombres portées et les variations importantes de l'éclairage. Afin de surmonter ces problèmes et d'obtenir de meilleurs résultats, plusieurs systèmes de détection de personnes, de suivi et de reconnaissance de posture humaine basés sur des caméras de profondeur, telles que la caméra stéréoscopique (Pellegrini and Iocchi, 2008; Satake and Miura, 2009), la caméra Kinect de Microsoft (Shotton et al., 2011; Choi et al., 2011; Le et al., 2013; J. Liu et al., 2013; Munaro et al., 2016; Liu et al., 2016) et la caméra à temps de vol (Time-of-Flight ou ToF) (Wientapper et al., 2009; Diraco et al., 2013; Luna et al., 2017) ont été proposés. Satake and Miura, (2009), par exemple, ont utilisé un ensemble de modèles de profondeur (depth templates) qui sont appliqués à des images denses de profondeur, puis ils détectent les personnes au moyen d'une technique de correspondance de modèles suivie par un vérificateur basé sur un SVM. Les résultats de la détection sont utilisés comme une entrée à un filtre de Kalman étendu pour effectuer le suivi de personnes. Choi et al., (2011) ont utilisé un ensemble de détecteurs multimodaux qui comprennent un détecteur de la partie supérieure du corps, un détecteur de visage, un détecteur de la peau, et un détecteur de forme et de mouvement basé sur la profondeur. Ces détecteurs sont ensuite combinés dans un cadre unifié en utilisant une méthode d'échantillonnage reposant sur le paradigme de suivi par détection. Liu et al., (2013) ont proposé de

détecter des personnes en utilisant un SVM entraîné avec deux caractéristiques différentes, à savoir l'histogramme de la partie supérieure du corps et l'histogramme conjoint de couleur et de hauteur de la tête humaine. Un filtre de Kalman et une technique simple d'association de données sont utilisés pour effectuer la tâche de suivi. Munaro et al., (2016) ont proposé une méthode de détection de personnes qui combine, en cascade, des classifieurs utilisant des caractéristiques de couleur et de profondeur. Le suivi est effectué à l'aide d'un filtre à particules qui exploite les histogrammes de couleur des personnes suivies. Liu et al., (2016) ont proposé de générer des cartes de vue en plan pour identifier des personnes candidates dans des régions spatiales d'intérêt, puis ils ont utilisé un histogramme de profondeur pondéré en combinaison avec un filtre particulaire pour suivre le mouvement des personnes détectées. Pour réaliser le suivi et la classification de posture humaine, Pellegrini and Iocchi, (2008) ont proposé un système qui repose sur la mise en correspondance (matching) de données 3D, provenant d'une caméra stéréoscopique, avec un modèle 3D du corps humain. Des points significatifs du modèle sont ensuite suivis à l'aide d'un ensemble de filtres de Kalman et, enfin, une classification basée sur le modèle de Markov caché est utilisée pour reconnaître la posture humaine au cours du temps. Shotton et al., (2011) ont traité le problème d'estimation de la posture comme une tâche d'étiquetage des parties du corps humain. Ce dernier est divisé en 31 parties différentes en fonction de l'emplacement de certaines articulations du squelette 3D, qui nécessitent d'être estimées. Un algorithme des forêts aléatoires est utilisé par les auteurs afin d'effectuer la classification des parties du corps. Le et al., (2013) ont proposé de reconnaître la posture humaine en utilisant les différents angles d'articulation du squelette du corps humain fournis par la caméra Kinect. Ces angles sont utilisés comme des vecteurs d'entrée à des SVM multi-classes pour reconnaître la posture humaine dans le cadre de surveillance et de suivi des soins de santé. Les systèmes basés sur des caméras de profondeur sont en général plus précis et plus résistants face aux changements d'éclairage et aux ombres portées. Cependant, la plupart de ces systèmes ne satisfont pas la contrainte du temps réel et leurs performances dépendent d'un grand nombre de paramètres, qui sont difficiles à régler. De plus, tout comme les systèmes basés sur des caméras visibles qui ne sont pas applicables dans certaines circonstances (par exemple, en cas de faible luminosité ou de nuit), les systèmes basés sur des caméras de profondeur sont limités

uniquement aux environnements intérieurs, tels que les chambres, les bureaux, les magasins, etc., et ce en raison du champ de vision restreint de ces types de caméras (par exemple, pour la caméra Kinect, son champ de vision est de 0,8 m à 4 m). Afin de pouvoir détecter et suivre des personnes quelles que soient les conditions d'éclairage (de jour ou de nuit) et quel que soit le type d'environnement (intérieur ou extérieur), il est nécessaire de recourir à d'autres solutions en utilisant d'autres types de caméras plus adaptées. Par ailleurs, au cours des dernières décennies, le coût des capteurs infrarouges a considérablement diminué, et des caméras infrarouges (IR), ayant une gamme dynamique et une sensibilité élevées, sont de plus en plus utilisées dans des applications telles que la vision nocturne et la surveillance en toutes circonstances. Motivés par cette baisse du coût des caméras IR, de nombreux travaux sur la détection et le suivi de personnes, et la reconnaissance de leur posture, dans des séquences d'images IR ont été récemment publiés (Olmeda et al., 2011; J. Wang et al., 2012; Qi et al., 2016; Yang et al., 2017; Soundrapandiyan and Chandra Mouli, 2018). Toutefois, ce sujet reste un problème difficile à gérer en raison de plusieurs facteurs. Parmi ces facteurs, on peut citer :

a) Les objets non-humains en mouvement tels que les animaux ou les véhicules.
b) Les encombrements de l'arrière-plan (background clutters), qui peuvent être provoqués par des objets ayant une forme ressemblant à celle d'un être humain, tels que des lampadaires, des troncs d'arbres, des arbustes, etc., et des mouvements aléatoires, tels que des branches d'arbres agités par du vent, des nuages ou des surfaces d'eau en mouvement pouvant facilement influencer les performances des algorithmes de détection et de suivi de personnes.
c) Les occultations d'une personne en mouvement par une autre personne en mouvement dans la scène ou par un objet fixe de l'arrière-plan. Dans ces deux cas, certaines parties de la personne peuvent être complètement ou partiellement cachées derrière d'autres personnes ou objets.
d) Les variations d'apparence que peuvent subir les personnes en mouvement en raison de certains facteurs, tels que la non-rigidité et la nature articulée du corps humain, les changements de point de vue de la caméra, les changements d'échelle, les occultations partielles et le bruit.

e) Le changement de posture d'une personne en mouvement qui peut se déplacer non seulement en posture "Debout" (Standing), mais aussi dans d'autres types de postures, tels que "Penché" (Bending), "Accroupi" (Squatting), ou "Rampant" (Crawling).
f) Le contraste et le rapport signal/bruit relativement faibles (Budzan and Wyżgolik, 2015) qui caractérisent les images infrarouges, ainsi que le manque d'informations (notamment la couleur) rendent plus difficiles les tâches de détection et de suivi de personnes en mouvement dans ce type d'images.
g) La contrainte du temps réel qui implique que le temps nécessaire entre le traitement de deux trames successives de la séquence d'images doit être suffisamment court pour que le système ne rate aucun événement important.

3. Nos contributions

Dans ce travail de thèse, nous présentons un nouveau système intelligent de vidéo surveillance capable de détecter et de suivre des personnes en mouvement, ainsi que de reconnaître leur posture, dans des séquences d'images infrarouges. Le schéma-bloc du fonctionnement de ce système est présenté dans la Figure 0.2.
Ainsi, nos principales contributions apportées dans ce système sont relatives à la distinction d'un être humain de tout autre objet en mouvement, son suivi et sa reconnaissance de posture.
Tout d'abord, pour distinguer un humain de tout autre objet, nous proposons deux différentes approches que nous appliquons juste après l'extraction des objets en mouvement à partir des séquences d'images en utilisant une méthode de soustraction d'arrière-plan basée sur un modèle de mélange de Gaussiennes (GMM). Ainsi, la première approche est basée sur le calcul d'une fonction de similarité combinée qui utilise des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement. Cette approche, malgré sa simplicité, est capable de détecter un être humain en mouvement dans des séquences d'images infrarouges sans la nécessité d'un apprentissage préalable (a priori) d'un modèle mathématique (classifieur). La seconde approche est basée sur la détection conjointe des deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre majuscule de l'alphabet grec Omega Ω), et les deux jambes. Cette approche, contrairement à la

plupart des méthodes d'état de l'art, a la capacité de détecter un être humain même en présence des changements dans sa posture.

Une fois qu'un humain en mouvement est détecté avec succès, afin de le suivre efficacement et de manière robuste en cas de présence de diverses situations réelles difficiles, telles que des occultations, des arrière-plans encombrés et de multiples humains en mouvement dans la scène surveillée, nous proposons une méthode qui utilise un filtre à particules et une combinaison adaptative d'informations provenant de plusieurs types de caractéristiques, à savoir, l'intensité, la texture, la vélocité de mouvement et la distance spatiale. Dans un premier temps, différents modèles dans ces différentes caractéristiques sont créés en ligne, puis sont ultérieurement mis à jour afin de les adapter aux changements significatifs de l'apparence de l'être humain détecté. Lorsque de nouvelles observations arrivent avec les trames suivantes, les distances de similarité entre les différents modèles créés et les régions en mouvement observées sont ensuite calculées. Ces distances de similarité individuelles sont enfin combinées, dans le cadre d'un filtre à particules, en utilisant des poids adaptatifs afin de suivre l'être humain détecté. Pour augmenter davantage la robustesse de notre méthode de suivi, nous introduisons aussi une stratégie automatique de détection et de traitement des occultations basée sur des règles heuristiques simples et l'histogramme de projection verticale (VPH) en niveaux de gris.

Parallèlement avec l'algorithme de suivi, nous effectuons l'opération de reconnaissance de posture de l'humain en mouvement. Pour décrire efficacement sa posture, nous extrayons de sa silhouette trois différentes caractéristiques, à savoir, les caractéristiques basées région, les caractéristiques basées contour et les caractéristiques géométriques. Ces caractéristiques sont tout d'abord combinées afin d'obtenir un seul vecteur de caractéristiques, qui est ensuite introduit dans un classifieur SVM multi-classes pour accomplir la tâche de reconnaissance de posture. Les résultats de reconnaissance de posture, conjointement avec les informations temporelles obtenues par l'algorithme de suivi, sont finalement exploités pour analyser le comportement de l'être humain détecté le long de la scène surveillée.

Les performances de notre système proposé sont évaluées sur des séquences d'images infrarouges capturées dans des environnements réels de nuit et les résultats expérimentaux obtenus ont démontré une bonne faisabilité et efficacité de

notre système pour la détection automatique des êtres humains en mouvement et l'analyse de leur comportement au cours du temps.

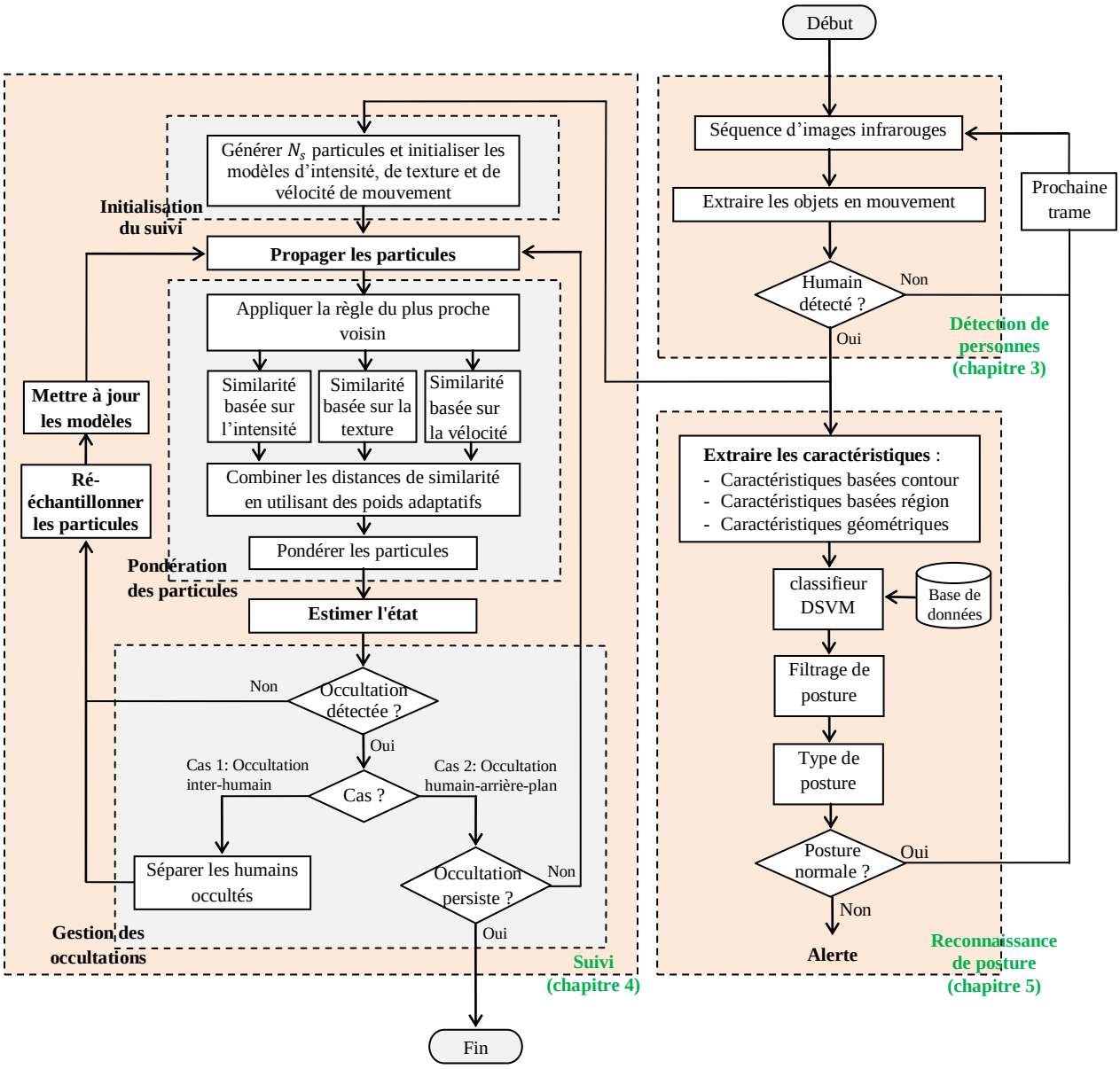


Figure 0.2 : Schéma-bloc de notre système proposé pour la détection, le suivi, et la reconnaissance de posture de personnes en mouvement dans des séquences d'images infrarouges.

4. Organisation de la thèse

Le reste de cette thèse est organisé comme suit :

Dans le Chapitre 1, nous rappelons quelques notions fondamentales nécessaires à la compréhension de la technologie de vision infrarouge. Initialement, nous présentons une brève définition du spectre électromagnétique et les différents domaines qui le constituent. Ensuite, nous présentons en plus de détails le domaine de l'infrarouge et les différents types de caméras utilisés pour la capture de ce type de rayonnement. Enfin, nous terminerons ce chapitre par la présentation de quelques applications des différentes sous-régions constituant le spectre infrarouge.

Dans le Chapitre 2, nous commençons, dans un premier temps, par présenter un état de l'art des principales méthodes développées pour la détection de personnes dans des séquences d'images infrarouges, puis nous présentons, dans un second temps, nos deux approches proposées dont l'une est basée sur le calcul d'une fonction de similarité qui combine des informations de divers types (forme, apparence, etc.), et l'autre, sur la détection conjointe de la partie tête-épaules et les deux membres inférieurs (jambes) du corps humain.

Dans le Chapitre 3, nous présentons d'abord un bref aperçu des techniques et des algorithmes les plus couramment utilisés pour le suivi d'objets en mouvement. Nous présentons ensuite en détail le principe du filtrage à particules, qui est l'une des techniques les plus largement utilisées dans le domaine du suivi d'objets. Enfin, nous terminerons ce chapitre par une présentation de notre approche de suivi, qui est basée sur l'algorithme du filtrage à particules et une combinaison adaptative d'informations provenant de plusieurs types de caractéristique.

Dans le Chapitre 4, nous décrivons notre nouvelle méthode de reconnaissance automatique de posture humaine que nous proposons. Celle-ci est basée sur trois caractéristiques de forme différentes, à savoir des caractéristiques basées région, des caractéristiques basées contour et des caractéristiques géométriques.

Dans le Chapitre 5, nous présentons les résultats expérimentaux obtenus par l'application de notre système sur des séquences d'images infrarouges réelles contenant différentes situations difficiles, telles que des occultations, des arrière-plans encombrés, des changements d'apparence, etc. Dans ce même chapitre, nous présentons aussi quelques résultats comparatifs avec d'autres travaux de la littérature afin de montrer les bonnes performances de notre système que nous proposons.

Dans la conclusion, nous rappelons les principales contributions de notre travail de thèse, puis nous discutons un certain nombre de limites du système proposé et nous décrivons quelques perspectives de recherche qui pourraient pallier ces limites.

Chapitre 1

Généralités sur la vision infrarouge

1.1. Introduction

Durant ces dernières années, la technologie de vision infrarouge est devenue de plus en plus efficace et ses domaines d'application sont en constante augmentation. Les améliorations techniques significatives des caméras infrarouges et les développements des techniques de traitement et d'analyse de l'image ont ouvert un large éventail de possibilités dans différents domaines, tels que l'agriculture, la médecine, l'industrie, l'astronomie, l'inspection, la sécurité et la surveillance.

Dans ce chapitre, nous présentons les notions fondamentales nécessaires à la compréhension de la technologie de la vision infrarouge. Tout d'abord, dans la Section 1.2, nous définirons brièvement le spectre électromagnétique et les différents domaines qui le constituent. Ensuite, dans la Section 1.3, nous présentons de manière détaillée le domaine de l'infrarouge étant donné que, dans cette thèse, nous traitons des images infrarouges ainsi que des différents types de caméras utilisées pour l'acquisition de ce type d'images. Dans cette même section, nous présentons aussi quelques applications des différentes sous-régions constituant le spectre infrarouge. Nous terminerons ce chapitre par une conclusion en Section 1.4.

1.2. Spectre électromagnétique

Le spectre électromagnétique constitue l'ensemble des rayonnements électromagnétiques qui s'étendent des rayons gamma aux ondes radios. Ces rayonnements, sous forme d'ondes électromagnétiques, sont caractérisés par deux propriétés fondamentales à savoir la fréquence f et la longueur d'onde λ. Il est connu que ces deux grandeurs sont liées par la relation suivante :

$$\lambda = \frac{c}{f} \tag{1.1}$$

où la vitesse de propagation de l'onde dans le vide c = 2.9979×10^8 m/s.

De cette expression, on rappelle que, plus la fréquence est élevée, plus la longueur d'onde est courte.

La représentation la plus simple du spectre électromagnétique est présentée dans la Figure 1.1. Ce spectre est découpé en plusieurs domaines disjoints selon leur fréquence ou leur longueur d'onde, comme le montre le Tableau 1.1.

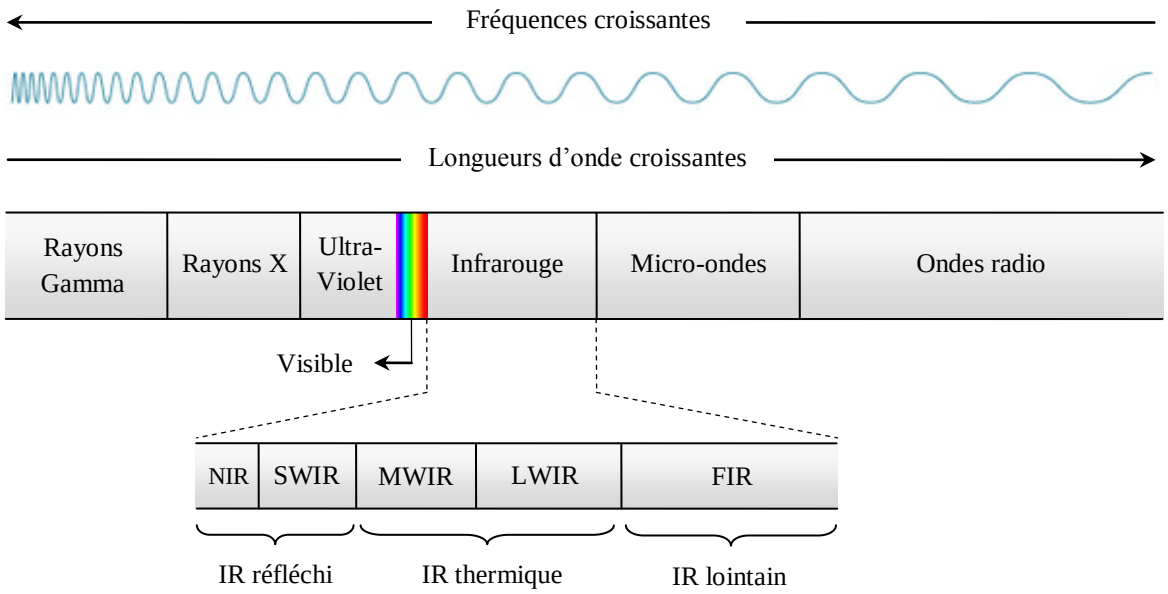


Figure 1.1 : Spectre électromagnétique.

Tableau 1.1 **:** Domaines du spectre électromagnétique.

Nom du domaine	Longueur d'onde [m]	Fréquence (Hz)
Rayons Gamma	$< 10\times10^{-12}$	$> 30\times10^{18}$
Rayons X	10×10^{-12} –10×10^{-9}	30×10^{15}–30×10^{18}
Ultraviolet	10×10^{-9}–400×10^{-9}	750×10^{12}–30×10^{15}
Visible	400×10^{-9}–700×10^{-9}	430×10^{12}–750×10^{12}
Infrarouge	700×10^{-9}–1×10^{-3}	300×10^{9}–430×10^{12}
Micro-ondes	1×10^{-3}–1×10^{0}	300×10^{6}–300×10^{9}
Ondes radio	1×10^{0} –100×10^{3}	3×10^{3}–300×10^{6}

C'est ainsi que nous distinguons selon les marges de ces longueurs d'onde, les rayons γ, les rayons X, les rayons ultraviolets, le spectre visible, l'infrarouge, les micro-ondes et les ondes radio.

Les rayons γ sont produits par la désintégration radioactive des noyaux atomiques lors de leur passage d'un état de haute énergie à un état inférieur appelé « décroissance gamma ». Leurs fréquences sont généralement supérieures à 30 exahertz (10^{18}), et de longueurs d'onde, donc, inférieures à 10 picomètres (10^{-12}), ce qui est inférieur au diamètre d'un atome. Les rayons gamma sont utilisés dans des domaines très variés tels que l'industrie (stérilisation et désinfection), la médecine (diagnostic, radiothérapie), et le nucléaire.

Les rayons X ont une longueur d'onde qui varie généralement de 0.01 à 10 nanomètres (10^{-9}), ce qui correspond à des fréquences de 30 petahertz (10^{15}) à 30 exahertz. Ces rayons trouvent de nombreuses applications, notamment dans le domaine de l'imagerie médicale (radiographies) et en astrophysique.

Les rayons ultraviolets, naturellement produits par le soleil et les étoiles chaudes, ou artificiellement par des lampes fluorescentes spéciales sont souvent subdivisés en trois catégories (Thieuleux et al., 2011) : l'ultraviolet long (UV-A), s'étendant de 400 à 315 nanomètres, l'ultraviolet moyen (UV-B), de longueur d'onde comprise entre 315 et 280 nanomètres, et l'ultraviolet court (UV-C), de longueur d'onde variant de 280 à 10 nanomètres.

Le spectre visible est la région du spectre électromagnétique que l'œil humain est capable de percevoir. Toutes les autres régions sont invisibles à l'œil humain sans matériel électronique spécial. Les rayonnements visibles sont produits par les vibrations et les rotations des atomes et des molécules, ainsi que par les transitions

électroniques au sein des atomes et des molécules. Il existe plusieurs définitions de la gamme des longueurs d'onde visibles. La limite inférieure est de l'ordre de 380 à 400 nanomètres (le violet), et la limite supérieure est de l'ordre de 700 à 780 nanomètres (le rouge) (Taillet et al., 2018). Sous des conditions artificielles, l'œil humain peut également percevoir des rayonnements électromagnétiques d'une longueur d'onde comprise entre 310 et 1 050 nanomètres (Dash and Dash, 2009).
Quant aux rayonnements infrarouges, ils ont des longueurs d'onde plus importantes que celles des rayonnements visibles. Dans la région du visible, le rouge est la couleur qui possède les plus grandes longueurs d'onde, donc les plus faibles fréquences. Les rayonnements infrarouges ont des fréquences inférieures à celles de la couleur rouge, d'où leur nom « infrarouge », du latin « infra » qui signifie « au-dessous » du rouge. Leurs longueurs d'onde sont généralement comprises entre 0.7 et 1000 micromètres (10^{-6}). Des détails supplémentaires sur la région de l'infrarouge sont décrits dans la Section 1.3.
En ce qui concerne les micro-ondes et les ondes radios, pour les premières, λ varie généralement de 0.001 à 1 mètre, et pour les deuxièmes, λ varie de 1 mètre à 100 kilomètres.
Notons que les limites entre les régions du spectre électromagnétique ne sont pas complètement rigides (Haken and Wolf, 2013), car des chevauchements entre des régions voisines peuvent également exister.

1.3. Le rayonnement infrarouge

Le rayonnement infrarouge, ou simplement l'infrarouge (IR), a été découvert en 1800 par l'astronome britannique d'origine allemande Sir Frederick William Herschel (1738-1822) (Rowan-Robinson, 2013) qui, en déplaçant un thermomètre dans le spectre visible obtenu en utilisant un prisme de verre, il observa une augmentation de la température des couleurs au fur et à mesure qu'elles passait du violet vers le rouge. Après avoir constaté cette découverte, il décida alors de mesurer la température juste au-delà de la couleur rouge du spectre dans une région apparemment dépourvue de lumière. À sa grande surprise, il découvrit que cette région avait une température plus élevée que celle de toutes les couleurs qui composent le spectre visible. Herschel a effectué d'autres expériences sur ce qu'il a appelé « rayons calorifiques » qui existaient au-delà de la couleur rouge du spectre visible et a découvert que ces rayons peuvent être réfléchis, réfractés, absorbés et

transmis tout comme la lumière visible. Ce qu'Herschel avait découvert était une forme de lumière au-delà du rouge. Ces « rayons calorifiques » ont ensuite été renommés, rayons (ou rayonnements) infrarouges.

Aujourd'hui, comme tout l'ensemble du spectre électromagnétique et certaines de ses régions, le spectre infrarouge est divisé en plusieurs domaines. Cependant, les frontières entre ces domaines ne sont pas universelles et elles peuvent légèrement varier suivant les domaines d'application. Selon la Commission Internationale de l'Eclairage (publication CIE No. 17.4 1987), le spectre infrarouge est divisé en trois grands domaines à savoir : IR-A, IR-B et IR-C. Les longueurs d'onde de ces domaines ainsi que leurs fréquences correspondantes sont montrées au Tableau 1.2. L'Organisation Internationale de Normalisation (norme ISO 20473:2007) divise aussi le spectre infrarouge en trois grands domaines qui sont montrés au Tableau 1.3. Dans ce chapitre, au lieu de ces deux divisions, nous suivrons la division montrée au Tableau 1.4, qui est la division la plus suivie au sein de la communauté scientifique (D'Amico et al., 2009; Gade and Moeslund, 2014; Picart, 2015). Selon cette division, le spectre infrarouge est constitué de 5 grandes régions : le proche infrarouge (NIR), l'infrarouge à courte longueur d'onde (SWIR), l'infrarouge à longueur d'onde moyenne (MWIR), l'infrarouge à longue longueur d'onde (LWIR), et l'infrarouge lointain (FIR). Le NIR et le SWIR combinés sont généralement appelés « infrarouges réfléchis », le MWIR et le LWIR combinés sont souvent appelés « infrarouges thermiques » ou « infrarouges émis ». Le FIR est parfois appelé « Terahertz », et il se situe à la limite avec le domaine des micro-ondes. De plus amples détails sur chacun de ces types d'infrarouges sont décrits dans les trois sous-sections suivantes.

Tableau 1.2 : Division du spectre infrarouge selon la Commission Internationale de l'Eclairage (publication CIE No. 17.4 1987).

Domaine	**Longueur d'onde**	**Fréquence**
IR-A	700–1400 nm	215–430 THz
IR-B	1400–3000 nm	100–215 THz
IR-C	3000 nm–1000 μm	300 GHz–100 THz

Tableau 1.3 : Division du spectre infrarouge selon l'Organisation Internationale de Normalisation (norme ISO 20473:2007).

Domaine	Abréviation	Longueur d'onde	Fréquence
Infrarouge proche	NIR (Near-InfraRed)	780–3000 nm	100–385 THz
Infrarouge moyen	MIR (Mid-InfraRed)	3000 nm–50 µm	6–100 THz
Infrarouge lointain	FIR (Far-InfraRed)	50–1000 µm	300 GHz–6 THz

Tableau 1.4 : Division du spectre infrarouge selon la communauté scientifique (D'Amico et al., 2009; Picart, 2015).

	Domaine	Abréviation	Longueur d'onde
Infrarouge réfléchi	Infrarouge proche	NIR (Near InfraRed)	0.75–1.4 µm
	Infrarouge court	SWIR (Short-Wavelength InfraRed)	1.4–3 µm
Infrarouge thermique (émis)	Infrarouge moyen	MWIR (Mid-Wavelength InfraRed)	3–8 µm
	Infrarouge long	LWIR (Long-Wavelength InfraRed)	8–15 µm
	Infrarouge lointain	FIR (Far InfraRed)	15–1000 µm

1.3.1. L'infrarouge réfléchi

L'infrarouge réfléchi couvre des longueurs d'onde allant d'environ 0.75 µm à 3 µm. Ce type d'infrarouge possède des propriétés physiques très similaires à celles de la lumière visible. Les photons incidents sont réfléchis ou absorbés par les objets de la scène, créant ainsi le contraste nécessaire pour générer des images utilisables. La lumière ambiante émise par la lune et les étoiles de l'univers est une source naturelle de l'infrarouge réfléchi qui offre un excellent éclairage dans des scènes extérieures de nuit. Cependant, dans les systèmes de vision nocturne, une source artificielle émettant des longueurs d'onde situant dans l'infrarouge proche (NIR) ou l'infrarouge court (SWIR) est souvent utilisée afin d'illuminer les scènes dans des conditions de faible luminosité. Les rayonnements réfléchis par les objets sont ensuite captés par une caméra sensible au NIR ou SWIR, qui les transforme

finalement en une image de la scène observée. Le schéma de principe de la vision par infrarouge réfléchit est montré dans la Figure 1.2.

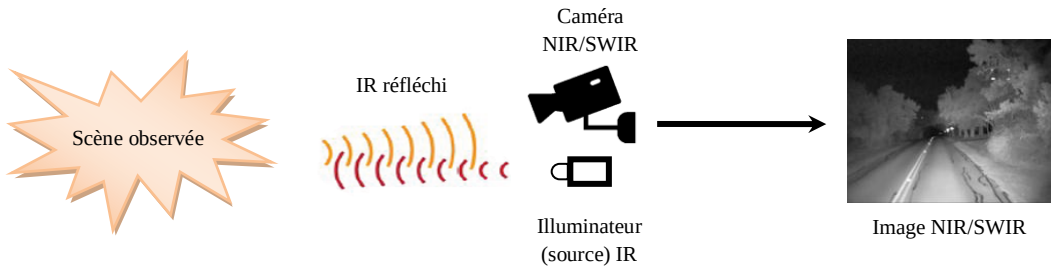


Figure 1.2: Schéma de principe de la vision par infrarouge réfléchi.

1.3.1.1. Types d'illuminateurs IR

Un illuminateur IR est une source artificielle de rayonnement infrarouge (invisibles à l'œil humain) qui permet à une caméra de capturer des images claires dans un environnement à faible visibilité ou complètement sombre. Il existe plusieurs types d'illuminateurs IR en fonction des besoins spécifiés. Ces illuminateurs reposent sur trois technologies différentes, à savoir les illuminateurs à incandescence, à LEDs et à LASER.

Les illuminateurs à incandescence sont fabriqués à partir des ampoules traditionnelles halogènes et au tungstène qui émettent de l'infrarouge réfléchi en plus d'une grande quantité de lumière visible très intense. Des filtres peuvent être utilisés pour permettre la projection uniquement de l'infrarouge sur la scène tout en réfléchissant la lumière visible vers le boîtier où elle est dissipée sous forme de chaleur. Les illuminateurs à incandescence sont moins chers à l'achat, mais leur durée de vie est généralement courte.

Les illuminateurs à LEDs utilisent un ensemble de diodes électroluminescentes infrarouges (InfraRed Light-Emitting Diodes ou IR-LEDs) à base de matériaux semi-conducteurs tels que l'Arséniure de Gallium (GaAs) ou l'Arséniure d'Indium et de Gallium (InGaAs). Ce type d'illuminateur produit beaucoup moins de chaleur que les illuminateurs à incandescence. Ils minimisent également le problème de la pollution lumineuse, car ils émettent une lumière IR invisible ou à peine visible. Ces illuminateurs sont souvent utilisés pour des éclairages à courte portée.

Les illuminateurs LASER sont basés sur le rayonnement LASER (Light Amplification by Stimulated Emission of Radiation), qui rappelons-le, est un

dispositif d'amplification de la lumière par émission stimulée de radiation. La lumière générée par ce dispositif est constamment stimulée par les photons générés. Cela signifie que lorsqu'un électron est illuminé par un photon, il se remplit d'énergie et s'élève à un niveau supérieur et, en revenant à son niveau initial, il émet un autre photon. Ce photon se réfléchit grâce aux miroirs à l'intérieur de la cavité du LASER puis illumine un deuxième électron. Le processus se répète en permanence et c'est ainsi qu'un LASER émet de la lumière. Les illuminateurs LASER sont excellents pour éclairer des scènes trop éloignées de la caméra, et même par mauvais temps. Cependant, en raison de leurs coûts très élevés, et leur dangerosité due à leur énergie et leur intensité élevées, les illuminateurs LASER ne sont utilisés que dans des applications militaires.
Les Figures 1.3 (a–c) suivantes montrent quelques exemples des différents types d'illuminateurs IR décrits ci-dessus.

(a) (b) (c)

Figure 1.3 : Exemples d'illuminateurs infrarouges. (a) Illuminateur à incandescence, (b) Illuminateur à LEDs, (c) Illuminateur LASER.

Les illuminateurs IR peuvent être aussi soit intégrés, attachables, ou portables. Dans le premier cas, les illuminateurs sont intégrés directement dans la caméra afin de lui fournir un éclairage précis dans son champ de vision. Ce sont les illuminateurs les plus communs dans les systèmes de vision nocturne infrarouge. Ils sont intéressants à courte distance. Dans le second cas, ils sont indépendants de la caméra et ils possèdent leur propre alimentation attachable. Ils ont tendance à être plus gros, plus lourds et plus volumineux que les illuminateurs intégrés. Ce type d'illuminateurs est souvent destiné à éclairer une zone spécifique de la scène observée. Les illuminateurs IR portables sont essentiellement des lampes de poche à vision nocturne. Ils sont souvent utilisés comme une source supplémentaire d'infrarouge dans les dispositifs de vision nocturnes.

Les Figures 1.4 (a–c) montrent des exemples d'illuminateurs IR intégrés, attachables et portables.

(a)

(b)

(c)

Figure 1.4 : Types d'illuminateurs IR. (a) Intégré, (b) Attachable, (c) Portable.

1.3.1.2. Les caméras à infrarouge réfléchi et leurs applications

Il existe deux types de caméras à infrarouge réfléchi : les caméras NIR et les caméras SWIR.

Les cameras NIR disponibles aujourd'hui utilisent des capteurs d'images à base de silicium (Si). Ces capteurs peuvent être soit de type CCD (Charge Coupled Device) ou CMOS (Complementary Metal Oxide Semi-conductor) (Teledyne DALSA, 2021). Les deux types de capteurs convertissent la lumière en charge électrique (électrons) proportionnelles à l'intensité lumineuse reçue. Dans un capteur CCD, la charge de chaque pixel est transférée à travers un nombre très limité (souvent un seul) de nœuds de sorties, puis elle est convertie en une tension qui est stockée et envoyée hors-puces sous forme d'un signal analogique. Comme tous les pixels peuvent être consacrés à la capture de la lumière, l'uniformité de la sortie est élevée, ce qui conduit à une meilleure qualité d'image. Dans un capteur CMOS, chaque pixel a son propre convertisseur charge/tension. Le capteur comprend souvent aussi des circuits supplémentaires pour l'amplification, la correction du bruit et la numérisation, ce qui permet à la puce de produire des bits numériques. Ces fonctionnalités supplémentaires augmentent la complexité de la conception et réduisent la surface disponible pour la capture de la lumière. Comme chaque pixel effectue sa propre conversion, il s'agit d'une architecture massivement parallèle ayant une largeur de bande totale élevée pour une vitesse élevée, mais l'uniformité de la sortie est plus faible. En raison de ces grandes différences, les capteurs CMOS sont plus adaptés aux dispositifs qui nécessitent de la vitesse et une faible consommation d'énergie, alors que les capteurs CCD sont plus adaptés pour les

applications qui nécessitent une haute qualité d'image et dans des conditions de faible luminosité.

Les capteurs CCD ou CMOS à base silicium utilisés dans les caméras NIR ne sont sensibles que jusqu'à 1100 nm, car le silicium devient transparent au-delà de cette longueur d'onde (Martin, 2015). Ainsi, les caméras SWIR exigent des composants optiques et électroniques spéciaux, capables de fonctionner dans la bande spectrale de 1400 nm à 3000 nm. Les détecteurs à base d'indium-arséniure de gallium (InGaAs), qui est un substrat semi-conducteur composé d'un mélange d'arséniure d'indium (InAs) et d'arséniure de gallium GaAs), ont été les premiers capteurs utilisés dans les caméras SWIR. Comme les capteurs à base de silicium, les capteurs en InGaAs sont des détecteurs photovoltaïques avec une jonction p-n, mais ils ont une énergie de bande interdite plus faible que le silicium, ce qui leur permet de détecter une plus grande gamme de longueurs d'onde. Les capteurs standards en InGaAs peuvent détecter des longueurs d'onde ayant de 900 nm à 1700 nm, et les capteurs en InGaAs étendus peuvent aller jusqu'à 2500 nm (Hamamatsu Photonics, 2021). Un avantage particulier des caméras SWIR à base de capteurs en InGaAs est leur capacité à générer des images de haute résolution et d'une qualité élevée avec un faible taux de bruit. De plus, ces caméras, contrairement à d'autres types de caméras SWIR, telles que celles à base de détecteurs en tellurure de mercure-cadmium (HgCdTe) ou en antimoniure d'indium (InSb), ne nécessitent ni d'obturateurs d'objectifs, ni de systèmes de refroidissement cryogéniques très coûteux. L'élimination de ces composants augmente la fiabilité des caméras, diminue leur coût, leur taille et leur poids, et les rend moins sensibles aux vibrations.

Comme application, les caméras à infrarouge réfléchi (NIR/SWIR) sont utilisées dans de nombreux domaines allant de la médecine à l'astronomie, tout en passant par l'agriculture et l'environnement, l'examen de documents, l'inspection et le contrôle de qualité.

En médecine, c'est dans l'imagerie médicale que les caméras à infrarouge réfléchi trouvent peut-être plus d'application. Cela revient à deux propriétés essentielles de l'infrarouge réfléchi. La première est sa capacité à pénétrer dans les couches superficielles de la peau et à révéler les structures situées en dessous. La deuxième est ses caractéristiques de réflexion et d'absorption qui diffèrent de celles du spectre visible. Ces deux propriétés constituent la base de toutes les applications médicales

des caméras à infrarouge réfléchi. Le sang veineux absorbe fortement les infrarouges, alors que le sang oxygéné les réfléchit très-bien. Ainsi, des troubles vasculaires tels que les varices ou l'obstruction veineuse sont clairement délimités. Les caméras à infrarouge réfléchi sont également utilisées pour enregistrer les variations du diamètre de la pupille dans de nombreuses études sur les effets physiques et physiologiques chez des sujets humains et animaux (comme l'œil n'est pas sensible à l'infrarouge, la pupille ne réagit pas à ce type de rayonnement). Dans les cas d'une opacité cornéenne, la taille de la pupille, sa forme et sa position peuvent être facilement déterminées à l'aide d'une caméra à infrarouge réfléchi. Ce type de caméra est aussi largement utilisé dans l'étude des tumeurs, notamment pour délimiter l'augmentation de l'apport sanguin aux tumeurs du sein, et pour différencier les lésions pigmentées bénignes et malignes de la peau.

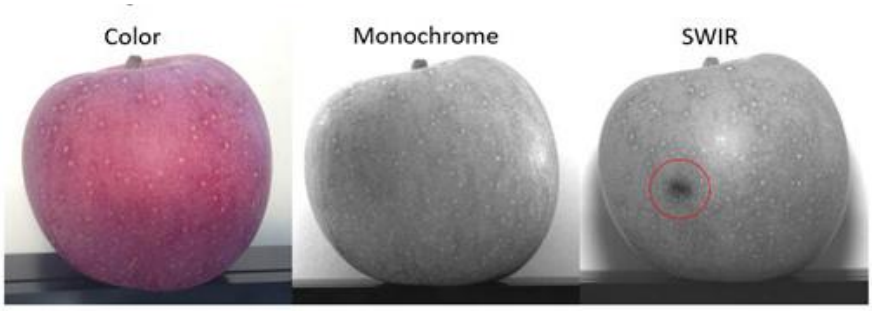
En agriculture et environnement, la propriété de réflexion et d'absorption de l'infrarouge réfléchi par les feuilles des arbres offre de nombreuses applications pour l'imagerie satellitaire et aérienne en botanique, en agriculture, en écologie, en foresterie et en aménagement du territoire. Différents types d'arbres et de cultures, ainsi que leur maladie, peuvent être facilement identifiés par une caméra à infrarouge réfléchi grâce à leur signature spectrale. L'infrarouge réfléchi est particulièrement très utile pour distinguer entre des eaux polluées et des eaux propres dans des images aériennes pour l'analyse du drainage et pour évaluer les dommages liés à la pollution par hydrocarbures. Les caméras à infrarouge réfléchi sont aussi utilisées dans l'étude des activités et des déplacements des populations animales dans leur milieu naturel.

Ce type de caméras peut également être utilisé pour diverses tâches d'inspection dans l'industrie alimentaire, l'industrie du bois, l'industrie textile ou l'industrie automobile ainsi que pour le contrôle qualité. Étant donné que l'eau est transparente à la lumière visible et qu'elle est très absorbante dans la gamme de longueurs d'onde infrarouges comprise entre 1 450 et 1 900 nanomètres (ce qui la fait apparaître plus sombre sur l'image), et en rajoutant certains filtres ou un éclairage approprié, cette capacité peut être utilisée dans différentes tâches telles que la détection sans contact du niveau de remplissage des conteneurs non transparents, la détection de fruits meurtris, moisis ou abîmés ainsi que la mesure de la teneur en humidité des produits secs en vrac (graines, riz, fruits secs, céréales, etc.).

Les caméras à infrarouge réfléchi trouvent aussi de nombreuses applications dans l'examen de documents, car elles permettent de révéler différents types d'encre, de découvrir des documents effacés ou réécrits, et de retrouver des écritures sur des documents noircis, vieillis ou brûlés. Comme l'infrarouge réfléchi peut pénétrer facilement dans des vieux vernis, ces caméras sont capables de révéler beaucoup d'informations supplémentaires sur les peintures, ainsi que d'établir leur chronologie, leur développement et de découvrir une fausse attribution.
En astronomie, l'infrarouge réfléchi, notamment l'infrarouge à courte longueur onde (SWIR), est très utilisé dans les études associées aux bandes photométriques J (entre 1µm et 1.4µm), H (entre 1.45µm et 1.8µm) et K (entre 2µm et 2.5µm). Ce type de rayonnement se distingue par plusieurs caractéristiques importantes parmi lesquelles son insensibilité aux phénomènes atmosphériques tels que les aérosols, la fumée, et le brouillard, ce qui permet d'obtenir des images de haute résolution et de très faibles niveaux de bruit.
La Figure 1.5 montre quelques exemples des applications des caméras à infrarouge réfléchi mentionnées ci-dessus.

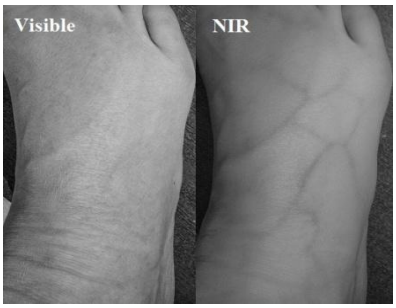
1.3.2. L'infrarouge thermique (ou émis)

Des « rayonnements thermiques » sont émis sous forme d'ondes électromagnétiques par tout objet possédant une température au-dessus du zéro absolu (0 K, –273.15°C). Ceux-ci diffèrent des autres types d'ondes (comme les ondes sonores) car elles ne nécessitent ni la présence d'un milieu, ni le déplacement de la matière. L'intensité de l'énergie émise varie en fonction de la température de l'objet, de la longueur d'onde du rayonnement, de l'air et du type de la surface de l'objet (Flir Systems, 2021). Un objet idéal, appelé « corps noir », ne réfléchit et ne transmet aucun rayonnement incident. Il absorbe la totalité de l'énergie qu'il reçoit, ce qui fait de lui un parfait absorbeur, et inversement, un parfait émetteur de radiations.

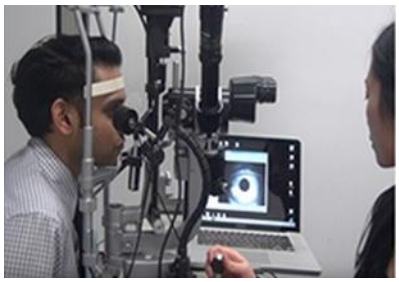


(a)

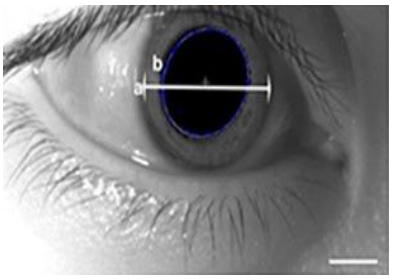
(b)



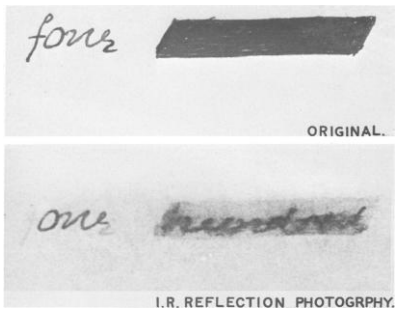
(c)



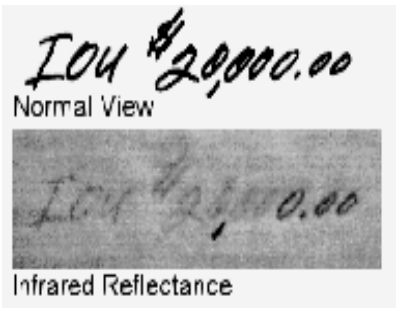
(d)



(e)



(f)



(g)

Figure 1.5 : Quelques exemples d'application des caméras à infrarouge réfléchi. (a) Détection de fruits abimés. (b) Détection du niveau de remplissage d'un conteneur non transparent. (c) Détection des veines. (d) et (e) Mesure du diamètre de la pupille. (f) et (g) Détection d'écriture oblitérée ou altérée.

Les propriétés radiatives d'un objet sont exprimées par son émittance globale ou émissivité. Cette émissivité globale, notée $\boldsymbol{\varepsilon}$, est égale au rapport de l'énergie émise par un corps radiant isotherme sur l'énergie émise par un corps noir à la même température. Il s'agit d'une mesure de la capacité d'un corps à rayonner l'énergie absorbée. Comme un corps noir est un parfait émetteur, la valeur de son émissivité est de 1. L'émissivité à une fréquence f est définie de manière similaire en considérant uniquement les radiations dont les fréquences sont comprises dans l'intervalle [f, f+df]. En équilibre thermique, l'énergie émise doit être équivalente à l'énergie absorbée par l'environnement du corps (Jones, 1998). La loi de Kirchoff stipule que pour tout corps uniforme en équilibre thermodynamique avec son milieu environnant, son émissivité est égale au rapport entre la puissance émissive de sa surface sur la puissance émissive d'un corps noir. Cela permet de définir l'émissivité $\boldsymbol{\varepsilon_f}$ d'un corps donné comme suite (Jones, 1998):

$$\boldsymbol{\varepsilon_f} = \frac{\boldsymbol{E_f}}{\boldsymbol{E_f^{noir}}} \qquad (\mathbf{1.2})$$

où, $\boldsymbol{E_f}$ est la puissance émissive de la surface d'un corps à une fréquence f, et $\boldsymbol{E_f^{noir}}$ est la puissance émissive d'un corps noir à la même température et longueur d'onde. Ainsi, l'émissivité $\boldsymbol{\varepsilon_f}$ est un nombre compris entre 0 et 1. Plus l'émissivité d'un objet est élevée, plus ses propriétés radiatives sont meilleures.

La loi de Planck du corps noir décrit les caractéristiques du rayonnement émis par un objet en termes de son émittance radiante spectrale, également connue sous le nom « puissance émissive ». En fonction de la longueur d'onde, la loi de Planck peut s'écrire comme suite (Grenn et al., 2012):

$$\boldsymbol{E_\lambda} = \varepsilon_\lambda \frac{2\pi hc^2}{\lambda^5} \frac{1}{\left(e^{\frac{hc}{\lambda k_B T}} - 1\right)} \qquad (1.3)$$

où :

- h = 6.6261×10^{-34} J s, est la constante de Planck.
- c = 2.9979×10^{8} m s^{-1}, est la célérité de la lumière dans le vide.
- k_B = 1.3807×10^{-23} W s K^{-1}, est la constante de Boltzmann.
- λ, est la longueur d'onde en mettre (m).

- T, est la température du corps en Kelvin (K).
- $\varepsilon_{\boldsymbol{\lambda}}$, est l'émissivité d'une surface à une longueur d'onde.
- $\boldsymbol{E_{\lambda}}$, est l'émittance radiante spectrale (ou puissance émissive) par unité de longueur d'onde et de surface [W m^{-2} m^{-1}].

Les courbes de la loi de Planck pour un corps noir à différentes températures sont données dans la Figure 1.6.

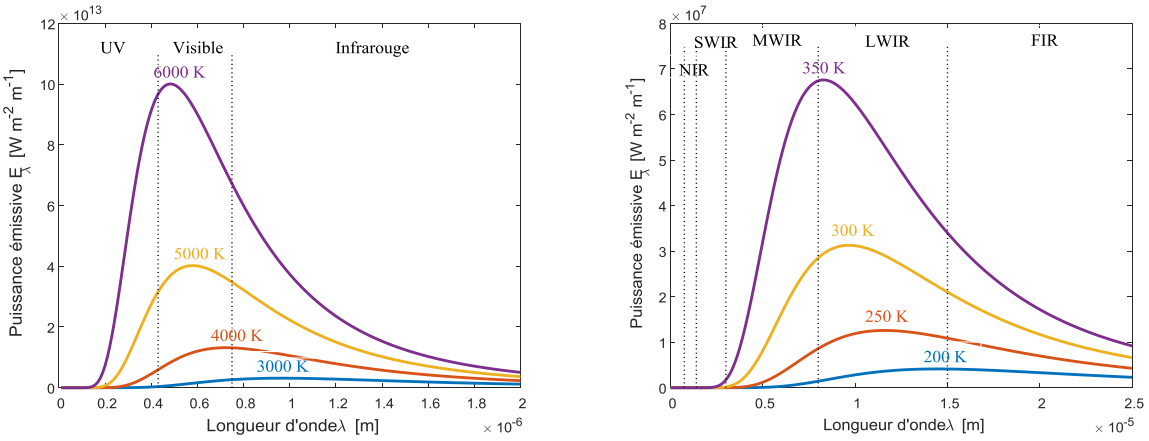


Figure 1.6 : Loi de Planck pour un corps noir à différentes températures.

Deux conclusions peuvent être tirées à partir des courbes de la Figure 1.6 : 1) la puissance émissive d'un objet augmente avec l'augmentation de sa température, et 2) la longueur d'onde λ_{max} correspondant au maximum de la puissance émissive $\boldsymbol{E_{\lambda}}$ se déplace vers les courtes longueurs d'onde au fur et à mesure que la température de l'objet augmente. Cette longueur d'onde pic peut être calculée en utilisant la loi du déplacement de Wien (Grenn et al., 2012), donnée dans l'équation (1.4), qui est obtenue en calculant la dérivée de la fonction de Planck par rapport à la longueur d'onde λ et en cherchant la valeur qui annule cette dérivée.

$$\lambda_{max} = \frac{b}{T} \tag{1.4}$$

Dans cette équation, b = 2.898×10^{-3} m K, est la constante du déplacement de Wien. Cette formule montre que le soleil, par exemple, ayant une température d'environ 5800 K, a un pic d'émission autour de 500 nm, qui se situe dans la gamme des longueurs d'onde visibles, alors que les objets de notre quotidien, à température ambiante (25°C, ou 298.15 K) et à température du corps humain (37°C, ou 310.15

K), émettent un pic de rayonnement autour de 10 µm, qui se situe dans la bande de LWIR (8–15 µm).

Comme un corps noir est un parfait absorbeur et un parfait émetteur, cela implique qu'il absorbe et émet le maximum d'énergie possible à une température donnée. La quantité de rayonnement ou puissance émissive totale, $\boldsymbol{E^{noir}}$ émise par un corps noir peut être également obtenue par la loi de Stefan-Boltzmann (Jones, 1998)**,** définie dans l'Equation (1.5), en intégrant la fonction de Planck sur toutes les longueurs d'onde (c.-à-d., de zéro à l'infini) :

$$E^{noir} = \int_{\lambda=0}^{\lambda=\infty} E_\lambda \, d\lambda = \sigma T^4 \quad (\text{en W m}^{-2}) \tag{1.5}$$

où, σ est la constante de Stefan-Boltzmann, qui est égale à:

$$\sigma = \frac{2\pi^5 k_B^4}{15h^3c^2} = 5.67 \times 10 - 8 \quad (\text{en W m}^{-2}\ \text{K}^{-4}) \tag{1.6}$$

L'équation (1.5) montre que la puissance émissive totale émise par un corps noir est proportionnelle à la quatrième puissance de sa température thermodynamique $\boldsymbol{T}$. Pour un corps non idéal (appelé corps gris) caractérisé par une émissivité $\varepsilon < 1$, la loi Stefan-Boltzmann, définie dans l'Equation (1.5), devient :

$$E^{gris} = \varepsilon\sigma T^4 \tag{1.7}$$

Cette relation entre l'énergie émise et la température est la base de l'imagerie infrarouge thermique. Elle montre que, lorsque l'émissivité d'un objet est connue, sa température peut être déterminée en mesurant son énergie rayonnée.

La Figure 1.7 montre le schéma de principe de la vision par infrarouge thermique (émis).

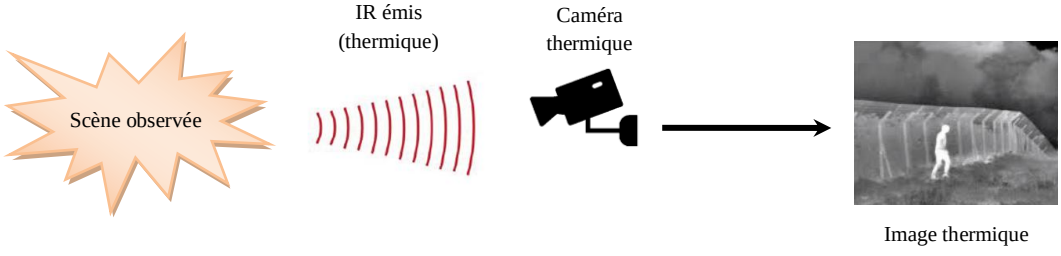


Figure 1.7: Schéma de principe de la vision par infrarouge thermique (émis).

1.3.2.1. Les caméras thermiques et leurs applications

Les caméras thermiques peuvent être divisées en deux grandes catégories selon le type de capteurs utilisés, à savoir caméras refroidies et caméras non refroidies.

Les caméras thermiques refroidies se composent de capteurs (ou détecteurs) d'images, de circuits électroniques d'interface et de traitement d'image, éventuellement d'une lentille de grossissement, et d'un refroidisseur cryogénique intégré. Le rôle de ce dernier est de réduire la température des capteurs à des valeurs cryogéniques, qui se situent autour de –200°C (75.15K) (Opgal, 2021). Cette réduction de la température des capteurs est nécessaire afin de réduire le bruit d'origine thermique à un niveau inférieur à celui du signal provenant de la scène observée. Si les capteurs (qui détectent et convertissent la lumière de la même manière que les caméras numériques conventionnelles, mais qui sont faits de matériaux différents) ne sont pas refroidis, ils risquent d'être envahis par leur propre rayonnement thermique, ce qui peut conduire à leur aveuglement. Les refroidisseurs cryogéniques, dont un exemple est montré dans la Figure 1.8, contiennent des pièces mobiles conçues à des tolérances mécaniques extrêmement serrées qui s'usent avec le temps. Ils contiennent également du gaz d'hélium (He), qui s'échappe lentement à travers les joints d'étanchéité. C'est pour ces raisons que ces refroidisseurs ont besoin d'un reconditionnement après 8 000 à 10 000 heures de fonctionnement (Flir Systems, 2018).

Figure 1.8 : Exemple de refroidisseur cryogénique (Flir Systems, 2018).

Les caméras thermiques refroidies sont les types de caméras les plus sensibles sur le marché aujourd'hui, et elles peuvent détecter des variations de température de l'ordre de 20 mK dans la scène observée. Elles sont généralement conçues pour capturer des images dans la bande de MWIR (3–8 µm) où, d'après la loi de Planck (Equation 1.3), le contraste thermique est élevé (Flir Systems, 2018). Rappelons ici,

que le contraste thermique est le changement relatif du signal pour un changement de la température de la cible. Plus ce contraste est élevé, plus il est facile de détecter des objets sur un arrière-plan qui peut être à une température très proche de celle de ces objets.

Les caméras thermiques à base des capteurs d'images refroidis sont plus coûteuses à fabriquer (donc plus chères à l'achat), et leur refroidissement peut parfois prendre plusieurs minutes avant d'être à bonne température pour être utilisables. Cependant, bien que les refroidisseurs cryogéniques soient encombrants et très coûteux, ils permettent aux caméras thermiques refroidies de capturer des images de très haute résolution et à une vitesse beaucoup plus élevée par rapport aux caméras non refroidies.

Quant aux caméras thermiques non refroidies, ce sont des caméras dont les capteurs d'images ne nécessitent pas de refroidissement cryogénique. La conception la plus courante des détecteurs de ces caméras est basée sur la technologie des micro-bolomètres. Ces derniers sont de minuscules résistances en oxyde de vanadium (Vox) ou en silicium amorphe (a-Si), présentant un coefficient de température élevé, placées sur une large surface en silicium ayant une faible capacité thermique et dotée d'une bonne isolation thermique (Flir Systems, 2018). Les changements de température de la scène observée provoquent des changements de température des bolomètres qui sont convertis en signaux électriques puis en une image thermique. Les capteurs non refroidis sont conçus pour fonctionner dans la bande de l'infrarouge à longue longueur d'onde LWIR (8–15 µm), où la plupart des objets terrestres émettent la majeure partie de leur énergie infrarouge.

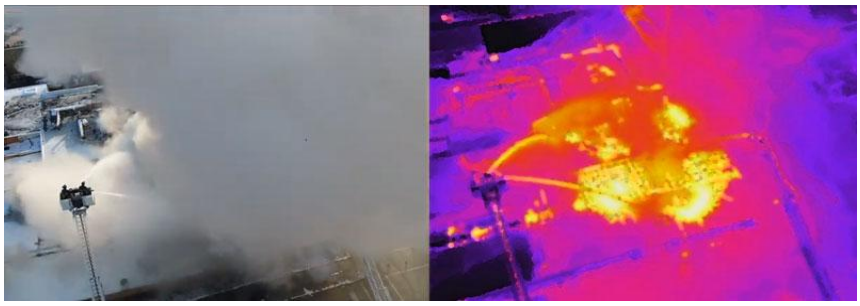
Les caméras thermiques non refroidies sont généralement beaucoup moins chères que leurs homologues, les caméras refroidies. La fabrication de leurs capteurs d'images nécessite moins d'étapes avec des rendements plus élevés par rapport aux capteurs refroidis, et leurs boitiers sous vide sont moins coûteux. Un autre avantage est que les caméras non refroidies ne nécessitent pas de cryo-refroidisseurs, qui sont des appareils très coûteux. Enfin, les caméras non refroidies comportent moins de pièces mobiles et ont tendance à avoir une durée de vie beaucoup plus longue que les caméras refroidies dans des conditions de fonctionnement similaires. Certaines applications telles que la sécurité et la surveillance nécessitent souvent un fonctionnement continu des caméras afin d'éviter de rater une menace potentielle. Les caméras refroidies doivent

généralement être entretenues après un à deux ans de fonctionnement, alors que les caméras non refroidies peuvent fonctionner en continu pendant des années.
Les caméras thermiques sont utilisées dans de nombreuses applications civiles ou de défense comme l'agriculture, l'inspection des bâtiments, l'industrie, la médecine, la médecine vétérinaire, la sûreté, la surveillance, et dans le domaine militaire.
Dans le secteur agricole, les caméras thermiques sont utilisées pour plusieurs tâches, notamment pour estimer le stress hydrique des sols et des cultures afin de planifier l'irrigation, mesurer le degré de salinité du sol dans les champs, déterminer les cultures infectées par des maladies et des agents pathogènes, cartographier la texture de surface des sols, estimer la quantité de résidus de culture laissés après la récolte, surveiller la maturité des cultures pour la récolte et cartographier le rendement des cultures. Les caméras thermiques sont également utilisées pour évaluer la viabilité des semences et des semis, détecter les dommages physiques et les désordres physiologiques sur les plantes, et surveiller le processus de croissance des semences, des semis et des plantes à l'intérieur des pépinières et des serres.
Dans l'inspection technique des bâtiments, l'infrarouge thermique est utilisé depuis des années et des caméras thermiques portables spéciales ont été développées en vue de cette application. Ces caméras permettent d'identifier plusieurs problèmes tels que la perte du flux de chaleur à travers les murs et les fenêtres, la présence d'une humidité excessive dans un bâtiment, et l'existence d'infiltrations ou de défauts d'étanchéité dans les systèmes de chauffage, de ventilation ou de climatisation. Les caméras thermiques permettent également d'évaluer les systèmes de toiture afin de détecter les fuites d'eau et les pertes d'isolation, qui peuvent entraîner des problèmes de moisissure dans les bâtiments, et qui sont susceptibles de mettre en danger la santé des personnes y travaillant ou y résidant.
Dans l'industrie, les caméras thermiques ont de nombreuses utilisations. Elles constituent un outil puissant et non invasif qui permet de surveiller et de dépanner rapidement, et avec précision, des équipements mécaniques et électriques tels que les pompes, les vannes de process, les réservoirs de stockage et les moteurs, afin d'assurer leur bon fonctionnement et d'éviter leur arrêt soudain et inattendu. Les caméras thermiques permettent également d'aider les techniciens à installer correctement les câblages et les équipements électriques afin d'éviter les blessures, les coûteuses coupures de courant et les dommages irréversibles.

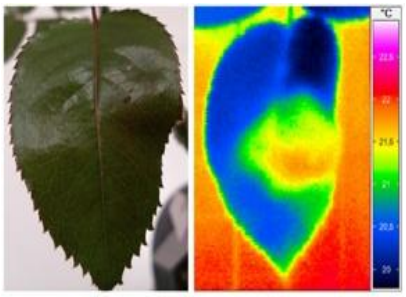
En médecine, les caméras infrarouges thermiques sont utilisées comme outil de diagnostic préliminaire dans un large éventail d'applications dans le domaine médical, telles que la détection du cancer du sein et de la peau, le dépistage de la fièvre, la surveillance des troubles thyroïdiens, l'identification des ulcères du pied chez le diabétique, l'identification des troubles de la circulation sanguine, le dépistage des problèmes du cou, du dos et d'épaule. En médecine vétérinaire, les caméras thermiques sont utilisées comme un moyen non invasif et efficace pour observer et examiner les animaux afin de dépister précisément et précocement certaines de leurs maladies et traumatismes. Elles sont également utilisées pour surveiller leur état physiologique, par exemple, pour mesurer l'efficacité de leur alimentation ou pour diagnostiquer une grossesse. Les caméras thermiques trouvent aussi des applications dans l'évaluation du bien-être des animaux, et elles sont couramment utilisées pour détecter les lésions chez les chevaux et surveiller leurs réactions au stress.

Dans le domaine militaire, les caméras thermiques sont utilisées dans une grande variété d'applications militaires pour les opérations de nuit ou en cas de visibilité réduite, ainsi que pour l'acquisition et la poursuite d'objectifs. Ces applications peuvent couvrir des environnements terrestres, aériens et maritimes. Des exemples d'applications comprennent la recherche et le sauvetage, la surveillance des cibles, la détection et l'évaluation des menaces, la détection des armes et des mines, et le guidage de précision des armes et des missiles. Les caméras thermiques sont aussi très utilisées dans les systèmes d'aéronefs télépilotés, ou drones.

Dans les domaines de la sécurité et la surveillance, les caméras thermiques trouvent de nombreuses applications telles que le contrôle des frontières, la sécurité intérieure et l'application de la loi. C'est un outil performant et très efficace qui permet de sécuriser des sites sensibles et des infrastructures critiques tels que les aéroports, les gares, les stations de métro, les centrales nucléaires et électriques, etc. Les caméras thermiques sont également utilisées par la police et les forces de l'ordre pour gérer les activités de surveillance, localiser et appréhender les suspects, enquêter sur les scènes de crime et mener des opérations de recherche et de sauvetage. Ces caméras sont aussi appliquées dans la lutte contre les incendies afin de détecter des personnes piégées ainsi que pour localiser la source d'un incendie.

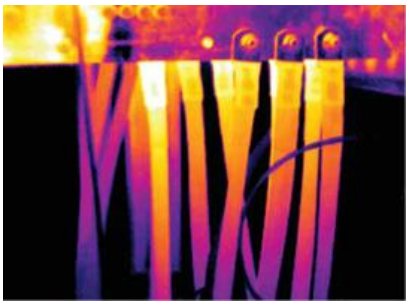


(a)

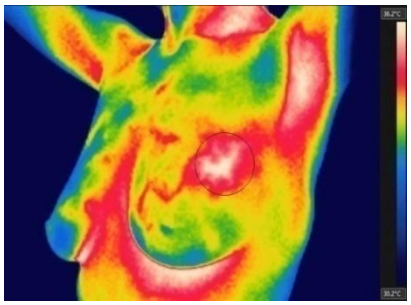


(b)

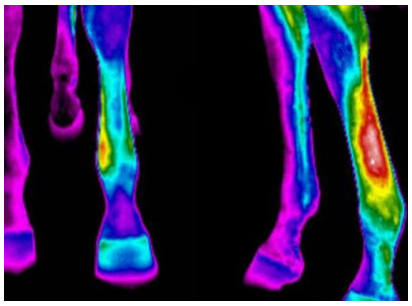
(c)



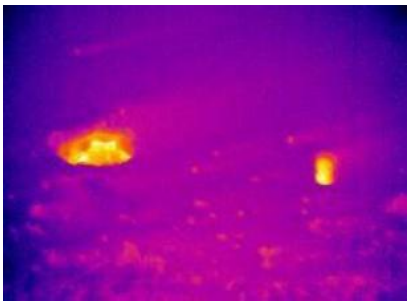
(d)



(e)



(f)



(g)

Figure 1.9 : Quelques exemples d'application des caméras thermiques. (a) Sûreté **:** localisation de la source d'incendie. (b) Agriculture : détection précoce des maladies. (c) Inspection des bâtiments : perte du flux de chaleur. (d) Industrie : localisation d'une surintensité du courant électrique. (e) Médecine : détection du cancer du sein**.** (f) Médecine vétérinaire : détection d'inflammation chez les chevaux. (g) Domaine militaire : détection des mines.

La Figure 1.9 donne quelques exemples d'applications des caméras thermiques mentionnées ci-dessus.

1.3.3. L'infrarouge lointain et ses applications

L'infrarouge lointain (Far Infrared ou FIR en anglais), appelé aussi la gamme des Terahertz (THz) (Picart, 2015; Shalaby et al., 2015), est la région du spectre électromagnétique contenant des longueurs d'onde ayant de 15 µm à 1000 µm, correspondant à des fréquences situées entre 300 GHz et 20 THz. Cela place le rayonnement infrarouge lointain entre la bande des infrarouges thermiques et celle des micro-ondes (voir Figure 1.1). D'autres sources utilisent d'autres limites pour l'infrarouge lointain. Par exemple, les astronomes définissent parfois l'infrarouge lointain comme les longueurs d'onde comprises entre 25 µm et 350 µm (Kim et al., 2017). Les rayons infrarouges lointains sont naturellement émis par les objets froids de l'univers (tels que les nuages moléculaires sombres qui engendrent de nouvelles étoiles et planètes) et ils sont détectés par des bolomètres (détecteurs de chaleur) refroidis à de très basses températures. De nos jours, les rayonnements infrarouges lointains sont générés par des lasers de plusieurs types, tels que les lasers au dioxyde de carbone (lasers au CO_2) (Patel, 2013) ou les lasers à électrons libres (Tan et al., 2012). A cause de leur excellente stabilité impulsion à impulsion et à long terme, et leurs systèmes d'exploitation relativement simples, les lasers femtosecondes (Cruz et al., 2007) sont les plus populaires pour les applications de l'infrarouge lointain. La majorité des travaux dans ce domaine ont été réalisés avec des lasers titane:saphir (également connus sous le nom de lasers Ti:Al_2O_3 ou Ti:saphir), fonctionnant à une longueur d'onde de 800 nm, qui est une longueur d'onde idéale pour piloter des émetteurs et des détecteurs de l'infrarouge lointain à base d'Arséniure de Gallium (GaAs).

Parmi les applications les plus importantes de l'infrarouge lointain en cours de développement, nous pouvons citer l'imagerie médicale, la sécurité, les communications, l'industrie et la recherche scientifique.

En imagerie médicale, sachant que les rayonnements infrarouges lointains sont capables de pénétrer profondément dans de nombreuses matières organiques sans causer de dommages associés aux rayonnements ionisants tels que les rayons X, ils peuvent être utilisés, entre autres, pour le diagnostic de certains cancers, tels que le cancer de la peau et le cancer du cerveau.

Les rayonnements infrarouges lointains peuvent être utilisés dans la surveillance, comme par exemple, pour le contrôle de sécurité, la détection d'armes dissimulées, et la détection de façon non destructive, et à distance, de stupéfiants ou de stimulants dans des courriers.
Dans le domaine des communications, les rayonnements infrarouges lointains trouvent plusieurs applications telles que les communications par satellites, les communications locales sans-fil, et les communications par radar.
Dans l'industrie, les rayonnements infrarouges lointains sont utilisés pour le contrôle de la qualité des procédés industriels et des produits de l'industrie plastique, la détection et la localisation des défauts de fabrication dans l'industrie aérospatiale, l'analyse et la surveillance des processus de revêtement, et la détection d'impuretés métalliques et non métalliques dans les produits de l'industrie agroalimentaire.
Dans le domaine de la recherche scientifique, les rayonnements infrarouges lointains sont utilisés dans plusieurs tâches telles que les mesures chimiques ou biochimiques, l'étude de la structure et des propriétés dynamiques de la matière condensée, la reconnaissance moléculaire, le repliement des protéines et l'astronomie submillimétrique.

1.4. Conclusion

Dans ce chapitre, nous avons présenté les principes fondamentaux pour la compréhension de la vision infrarouge. Nous avons tout d'abord présenté brièvement la définition du spectre électromagnétique et les différentes régions qui le composent. Nous avons ensuite défini en détail la région de l'infrarouge et les différents types de caméras utilisés pour la capture de ce type de rayonnement. Nous avons également présenté quelques exemples d'application des différentes sous-régions constituant le spectre infrarouge.
Dans le prochain chapitre, nous présenterons nos deux approches proposées pour la détection de personnes dans des séquences d'images IR. La première de ces méthodes est basée sur le calcul d'une fonction de similarité qui combine des informations de divers types (de forme, d'apparence, etc.), alors que la deuxième est basée sur la détection conjointe de la partie tête-épaules et les deux membres inférieurs (jambes) du corps humain.

Chapitre 2

Détection de personnes dans des séquences d'images IR

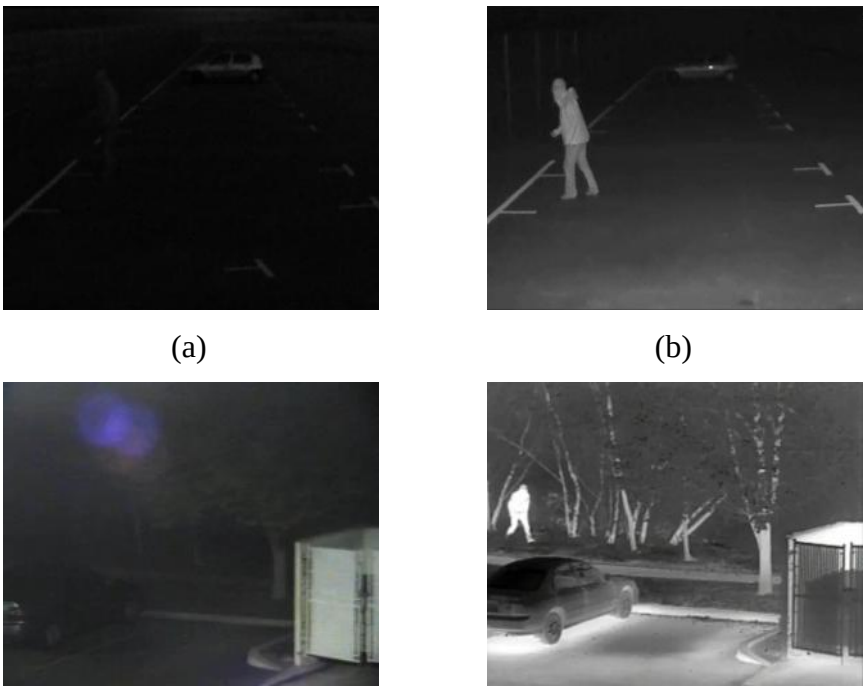
2.1. Introduction

La détection automatique de personnes, qui consiste à localiser toutes les formes de corps humains présentes dans des images ou des séquences d'images avec la plus grande précision possible (Davis et al., 2009), constitue un domaine de recherche très actif en vision par ordinateur, car elle représente une étape très importante pour de nombreuses applications, telles que la vidéosurveillance intelligente, les systèmes d'aide à la conduite, la recherche d'images/vidéos par le contenu, la robotique et les applications militaires. De plus, en raison des progrès technologiques remarquables effectués au cours de ces dernières années dans le but d'optimiser la puissance de traitement des ordinateurs, la détection de personnes a fait l'objet d'une attention considérable de la part de la communauté de vision par ordinateur, et un grand nombre de méthodes et de techniques ont été proposées, dont la plupart sont destinées pour des caméras standards fonctionnant dans le domaine du visible. Parmi les méthodes les plus populaires dans ce contexte, nous pouvons citer les caractéristiques pseudo-Haar (Viola et al., 2005; Guo et al., 2012), le SIFT (Scale-Invariant Feature Transform) (Mikolajczyk et al., 2004; Seemann et al., 2005), le SURF (Speed-Up Robust Features) (Bay et al., 2008), l'Histogramme d'Orientation du Gradient (HOG) (Dalal and Triggs, 2005; Pang et al., 2011), les motifs binaires locaux (Local Binary Patterns, LBP) (Yadong Mu et al., 2008; Satpathy et al., 2013), et la méthode Local Self-Similarity (LSS)

(Shechtman and Irani, 2007). Afin d'obtenir de meilleurs résultats, d'autres approches dites hybrides (Abari, 2018; Yao et al., 2015a, 2015b) combinent ces différentes méthodes les unes avec les autres. Pour un aperçu plus complet des travaux récents sur la détection de personnes dans des images acquises par des caméras fonctionnant dans le domaine visible, le lecteur est invité à se référer aux travaux de (Nguyen et al., 2016), et de (Bali and Tyagi, 2018). Bien que les méthodes de détection de personnes basées sur des caméras visibles aient des performances tout à fait acceptables, leur efficacité reste toutefois limitée notamment en cas de présence dans la scène de certains facteurs tels qu'un éclairage non uniforme, les ombres portées, et une faible luminosité extérieure (durant le soir et la nuit). Pour pallier ces différents problèmes, la détection de personnes en utilisant des caméras infrarouges, qui sont capables de capturer des images dans des conditions non conventionnelles (une obscurité totale, par exemple) est considérée comme une alternative.

Ainsi, l'utilisation des caméras infrarouges a été historiquement limitée aux applications militaires, sécuritaires et médicales. Cependant, avec le développement des technologies de fabrication des capteurs infrarouges au cours de ces dernières années, le coût des caméras infrarouges a considérablement diminué. En conséquence, de nombreux systèmes pour la détection de personnes dans des images et séquences d'images acquises par des caméras infrarouges ont été proposés dans la littérature. En fonction du type de caméra infrarouge utilisée, ces systèmes peuvent être généralement regroupés en deux grandes catégories : les systèmes basés sur des caméras proche IR et les systèmes basés sur des caméras IR thermiques (J. H. Lee et al., 2015; Y. Lee et al., 2015; Piniarski and Pawłowski, 2016). Les systèmes de la première catégorie illuminent activement la scène dans le spectre du proche infrarouge et capturent le rayonnement réfléchi par les objets présents dans cette même scène (y compris des êtres humains), alors que les systèmes de la deuxième catégorie génèrent des images en détectant passivement les émissions thermiques émises par les objets de la scène. Chacun de ces types de systèmes a ses propres avantages. Les systèmes basés sur des caméras proche IR, par exemple, se caractérisent par une résolution d'image plus élevée, et ils fournissent des images faciles à interpréter pour les utilisateurs en raison de la proximité du spectre proche IR au spectre visible (Li et al., 2017). Quant aux systèmes basés sur des caméras IR thermique, ils ne nécessitent aucune source

d'illumination infrarouge et les personnes apparaissent comme des objets lumineux dans la scène. La Figure 2.1 illustre deux exemples d'images infrarouges (proche IR et IR thermique) comparées avec des images visibles prises pour la même scène et dans les mêmes conditions (faible luminosité). Cette figure montre l'avantage et la capacité des systèmes basés sur des caméras infrarouges, par rapport aux systèmes basés sur des caméras visibles, à percevoir et à détecter des personnes en mouvement dans des environnements contenant des situations difficiles, telles qu'un éclairage faible.



(a) (b)

(c) (d)

Figure 2.1: Des personnes en mouvement dans des environnements extérieurs à faible luminosité. Les personnes ne sont pas perceptibles dans les images visibles (a) et (c), mais elles sont facilement détectables dans (b) les images proche IR et (d) les images thermiques.

Cependant, bien que de nombreux systèmes aient été proposés au sein de la communauté de vision par ordinateur, la tâche de détection de personnes à partir des images et des séquences d'images acquises par des caméras IR reste un problème difficile à traiter en raison des différents défis décrits précédemment dans l'introduction, à savoir la présence d'encombrements (clutters) et d'objets non humains en mouvement dans l'arrière-plan, les changements de posture et

d'apparence, les occlusions, un faible contraste et rapport signal/bruit, et la contrainte du temps réel. Pour relever ces différents défis, dans ce chapitre, nous proposons deux approches différentes pour la détection automatique de personnes en mouvement dans des séquences d'images IR. La première approche est basée sur le calcul d'une fonction de similarité combinée qui utilise des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement. Cette approche peut être appliquée sur des séquences d'images acquises par des caméras IR thermiques, et malgré sa simplicité, elle a comme avantage de détecter un être humain en mouvement sans la nécessité d'un apprentissage préalable d'un classifieur. La deuxième approche que nous proposons est basée sur la détection conjointe des deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre majuscule de l'alphabet grec Omega Ω), et les deux jambes. Cette approche, comme elle est basée sur la silhouette des objets en mouvement, elle peut être appliquée sur des séquences d'images acquises par des caméras thermiques ou par des caméras proche IR. L'avantage principal de cette approche est que, contrairement à la plupart des méthodes d'état de l'art, que nous exposerons dans la section suivante, a la capacité de détecter un être humain même en cas de présence de changements de sa posture.

2.2. Etat de l'art sur la détection de personnes dans une séquence d'images IR

Selon le type de caméras utilisées, les méthodes de détection automatique de personnes dans des images et des séquences d'images infrarouges peuvent être classées en deux grandes catégories : les méthodes basées sur des caméras proche IR, et les méthodes basées sur des caméras IR thermiques. Dans cette section, nous proposons de passer en revue les principales méthodes proposées dans la littérature dans chacune de ces catégories. Ces méthodes, en général, se composent de deux étapes principales : une étape de génération des régions d'intérêt (Region-Of-Interest, ROI) dans laquelle les régions qui peuvent potentiellement contenir des personnes sont extraites des images IR, et une étape de validation dans laquelle des caractéristiques sont extraites à partir des ROIs, puis elles sont introduites à un classifieur qui réalisera la détection.

2.2.1. Méthodes basées sur des caméras proche IR

Jusqu'à présent, différentes méthodes ont été proposées pour résoudre le problème de détection de personnes à partir des images et des séquences d'images acquises par des caméras proche IR. Andreone et al., (2005) ont proposé d'utiliser la transformée en ondelettes de Haar, avec trois fonctions de base différentes (horizontale, verticale, et diagonale), pour extraire des caractéristiques à partir des ROIs dans les images IR, puis ils sont utilisés un classifieur à base de SVM pour différencier entre un humain et un objet non-humain. Dong et al., (2007) ont proposé une méthode qui est composée de trois étapes principales. Une première étape, appelée étape de sélection, qui consiste à séparer les objets d'intérêt de l'arrière-plan de la scène en utilisant un algorithme de segmentation adaptatif à double seuil. Une deuxième étape, appelée étape de prétraitement, qui consiste à rejeter la grande majorité des objets non humains (sélectionnés dans la première étape) en appliquant un certain nombre de contraintes géométriques et spatiales sur les objets d'intérêt. Enfin, une dernière étape, appelée étape de reconnaissance, qui consiste à reconnaitre des personnes en utilisant une cascade de classifieurs basée sur les Histogrammes d'Orientation du Gradient (HOG) et l'algorithme AdaBoost (Adaptive Boosting). Soga et al., (2008) ont utilisé une cascade boostée de classifieurs pour extraire des régions candidates (régions ayant une intensité lumineuse élevée dans l'image), puis ils ont proposé de vérifier ces régions à l'aide d'une méthode de vérification qui utilise un SVM et quatre caractéristiques directionnelles différentes extraites en appliquant l'opérateur de Prewitt. En tenant compte à la fois de la faible complexité de calcul des caractéristiques pseudo-Haar et de la bonne capacité discriminative du descripteur HOG, Ge et al., (2009) ont proposé un système de détection de personnes dans des images proche IR en utilisant une cascade de classifieurs à base d'arbres de décision entraînés à l'aide de l'algorithme AdaBoost. Les caractéristiques pseudo-Haar sont utilisées pour vérifier rapidement la présence d'un corps humain au sein des régions d'intérêt, extraites en utilisant un algorithme de segmentation à double seuil, alors que le descripteur HOG est utilisé dans une deuxième phase de détection pour effectuer une vérification plus approfondie et pour s'assurer de la présence d'une personne. Lin et al., (2011) ont combiné deux caractéristiques importantes, à savoir le HOG et le contour, puis ils ont utilisé l'algorithme SVM pour établir un système de classification fiable. Les auteurs ont également proposé une technique de

segmentation intelligente pour l'extraction des régions candidates à partir de l'arrière-plan. Kancharla et al., (2011) ont proposé une méthode de détection de personnes à partir des images proche IR en utilisant des caractéristiques de bords extraites à l'aide de trois noyaux de Sobel (un vertical et deux diagonaux). Les régions candidates sont sélectionnées par des algorithmes de détection et de fusion de blobs. Un algorithme de mise en correspondance de modèles (template matching) et des règles heuristiques, basées sur des critères de taille et de forme, sont enfin appliquées sur les régions candidates afin de réduire le nombre de fausses détections. Zin et al., (2011) dans leur travail ont proposé une méthode hybride, dans laquelle des caractéristiques multi-fentes (multi-slits features) et des HOG sont fusionnées afin de détecter des personnes à partir des images proche IR. Cette fusion permet de tirer profit des avantages des deux caractéristiques pour l'identification et la localisation des différentes parties du corps humain, à savoir la tête, le torse et les jambes. Kumar, (2013) a utilisé le filtre de Gabor pour extraire les bords verticaux potentiels dans l'image proche IR, puis il a utilisé leurs profiles vertical et horizontal pour obtenir les boîtes englobantes des régions candidates. Ces régions sont ensuite soumises à une série de classifieurs utilisant différentes caractéristiques (la hauteur, la texture, le contraste, etc.) afin d'éliminer les fausses positives. Afin d'obtenir des meilleurs résultats, Govardhan and Pati, (2014) ont proposé une approche similaire à celle de Ge et al., (2009) en combinant les caractéristiques de Haar et le HOG. Dans une première étape, une cascade basée sur les caractéristiques de Haar est utilisée afin de détecter le corps humain en entier. Cela permet d'éliminer un grand nombre de régions non-humaines présentes dans l'image proche IR. Ensuite, afin de raffiner les résultats de détection, une deuxième étape de détection qui consiste à détecter les parties supérieure et inférieure du corps humain en utilisant un SVM et le descripteur HOG est effectuée. Un système de validation du corps entier est également mis en œuvre lorsqu'une des détections des deux parties du corps humain échoue. Pour résoudre le problème des occultations partielles dues à l'absorption des rayons IR par les vêtements portés par les personnes, Lee et al., (2015) ont proposé une méthode de détection basée sur l'appariement (matching) et le regroupement des modèles de parties. Trois parties du corps humain, à savoir la partie tête-épaules, le torse et les jambes sont tout d'abord identifiées individuellement en utilisant certaines contraintes et les relations spatiales entre chaque paire de ces parties. Le résultat

global de la détection de personnes est ensuite affiné par une méthode de segmentation par blocs. Han and Song, (2016) ont proposé une méthode de détection de personnes dans des images proche IR en utilisant des caractéristiques à canaux agrégés (Aggregated Channel Features, ACF) et l'algorithme AdaBoost. Une méthode de prétraitement adaptative basée sur une représentation en couches des différences de niveaux de gris et l'algorithme d'Otsu est également proposée par les auteurs pour améliorer le contraste des personnes. Li et al., (2017) ont proposé une méthode de détection de personnes en utilisant un système de vision stéréoscopique composé de deux caméras proche IR. Un modèle tridimensionnel de surface de voxel (Three-Dimentional Voxel Surface Model) est utilisé par les auteurs pour supprimer les pixels de l'arrière-plan et segmenter, en temps réel, les régions d'avant-plan représentant les personnes en mouvement dans la scène. Plus récemment, Dai et al., (2019) ont proposé une méthode de détection de personnes en combinant un réseau de neurones convolutifs (Convolutional Neural Network, CNN) et des images acquises à l'aide d'une caméra proche IR montée sur une voiture. Un modèle de CNN auto-apprenant à 9 couches basé sur la fonction d'activation soft-max est utilisé par les auteurs dans cette méthode, et les résultats obtenus sur un ensemble de données constitué de 15000 échantillons de tests ont été satisfaisants.

2.2.2. Méthodes basées sur des caméras IR thermiques

Un nombre considérable de travaux de recherche ont été publiés au cours des dernières décennies pour détecter des personnes à partir des images et des séquences d'images acquises par des caméras IR thermiques. Un état de l'art et une revue critique de certaines de ces méthodes est présenté dans (Negied et al., 2015). Ces méthodes, comme nous l'avons mentionné précédemment, se composent en général d'une étape de génération des ROIs et d'une étape de validation.

L'une des approches les plus populaires pour la génération des ROIs, mais aussi parmi les plus coûteuses en termes de temps de calcul, est l'approche de la fenêtre glissante (Qi et al., 2016), qui consiste à balayer, de manière exhaustive, l'ensemble de l'image d'entrée en utilisant des fenêtres de recherche à des positions et des tailles variables. Inspirés par cette approche, Sun et al., (2011) ont présenté une technique de génération des ROIs en utilisant des fenêtres glissantes locales centrées sur des points d'intérêt. Dans cette méthode, tous les points d'intérêt

candidats dans les images thermiques sont détectés en utilisant le détecteur SUSAN (Smallest Univalue Segment Assimilating Nucleus) (Smith and Brady, 1997), puis les ROIs sont générées par l'extraction des sous-fenêtres dans le voisinage de ces points d'intérêt détectés. Cependant, due à leur température corporelle relativement élevée, les humains apparaissent souvent comme des objets à intensité lumineuse élevée sur les images thermiques, et cette caractéristique particulière n'a pas été exploitée par Sun et al., (2011). Bertozzi et al., (2007) ont proposé de générer les ROIs en utilisant une méthode de segmentation d'images basée sur un seuillage global. Le seuil a été défini à partir des propriétés statistiques des images thermiques pré-collectées contenant uniquement des objets de l'arrière-plan, et ce, parce que les pixels des personnes dans la base de données utilisée par les auteurs étaient d'une intensité lumineuse élevée par rapport à ceux des objets de l'arrière-plan. Cependant, la méthode de segmentation par un seuillage global est difficile à utiliser à cause des changements d'apparence des personnes, notamment dans les environnements extérieurs non contrôlés. A la différence de la méthode conventionnelle de croissance de régions utilisée par Chen et al., (2008), les ROIs peuvent être générées en utilisant une méthode de croissance de régions basée sur des caractéristiques. Les germes sont souvent les pixels correspondant aux régions chaudes ayant une haute intensité dans les images IR thermiques (O'Malley et al., 2010). L'algorithme de croissance s'arrête lorsque les boîtes englobantes des régions connexes ne couvrent plus les intervalles possibles des rapports d'aspect des personnes. Une approche alternative pour l'extraction des ROIs est proposée par Yajun Fang et al., (2004). Dans cette approche, la position horizontale des ROIs est estimée en utilisant une projection horizontale basée sur l'intensité des pixels, alors que leur position verticale est estimée par une segmentation verticale basée sur l'intensité et les deux lignes extrêmes (gauche et droite) du corps humain. Li et al., (2010) ont proposé une approche similaire en combinant les projections horizontales et verticales basées sur l'intensité. L'avantage de ces deux approches réside dans leur flexibilité à s'adapter aux différents environnements. Cependant, la précision des ROIs générées dépend fortement de la qualité des images infrarouges, car elles supposent que l'intensité des pixels des personnes soit supérieure à l'intensité moyenne des pixels de l'arrière-plan de la scène. Au lieu d'utiliser les projections horizontales et verticales en se basant sur l'intensité, Q. Liu et al., (2013) ont proposé de segmenter les images thermiques en plusieurs bandes

verticales en utilisant la courbe de projection verticale du gradient des pixels, puis ils ont appliqué la méthode de segmentation à double seuillage de Ge et al., (2009) à l'intérieur de ces bandes afin de générer les ROIs.

Après avoir généré un ensemble de ROIs, l'étape suivante de validation, qui consiste à détecter des personnes à partir de ces ROIs, peut être effectuée. Pour cela, des algorithmes et des techniques d'apprentissage machine (machine learning) exploitant des descripteurs pour représenter et décrire les personnes sont souvent utilisés. Parmi les descripteurs les plus couramment employés, nous pouvons en citer : les caractéristiques pseudo-Haar (Benezeth et al., 2008; Qi et al., 2016), le HOG et ses variantes (Suard et al., 2006; Liu et al., 2013; O'Malley et al., 2010; Qi et al., 2016), l'histogramme de congruence de phase (Histogram of Phase Congruency, HPC) (Olmeda et al., 2012; Qi et al., 2016), les caractéristiques à canaux agrégés (Aggregated Channel Features, ACF) (Brehar et al., 2014; Kim and Kim, 2018), les motifs binaires locaux (LBP) et leurs variantes (Dong Xia et al., 2010; Y. Liu et al., 2017; Sun et al., 2011). D'autres descripteurs incluent l'auto-similarité d'intensité (Intensity Self Similarity, ISS) (Miron et al., 2012), les edgelets (Li Zhang et al., 2007), l'histogramme de codes épars (Histogram of Sparse Codes, HSC) (Qi et al., 2016), la transformée en ondelettes (Li et al., 2010), la transformée en cosinus discrète (Teutsch et al., 2014), l'histogramme de distribution de forme (Shape Distribution Histogram, SDH) (Zhao et al., 2015), et les moments discrets de Chebyshev (Lahouli et al., 2018). Une fois que les caractéristiques sont extraites en utilisant des descripteurs, des algorithmes d'apprentissage machine, tels que les SVMs (J. Wang et al., 2012; Y. Liu et al., 2017; Lahouli et al., 2018), l'AdaBoost (Kim and Kim, 2018; Y. Liu et al., 2017), la classification Bayésienne (Nanda and Davis, 2002; Teutsch et al., 2014), la classification par une représentation éparse (Sparse Representation Classification, SRC) (Zhao et al., 2015; Qi et al., 2016), et les réseaux de neurones à convolution (CNNs) (John et al., 2015; Herrmann et al., 2016; Heo et al., 2018; Park et al., 2020) sont utilisés afin de procéder à la classification des ROIs comme humain ou non-humain.

L'un des premiers travaux pionniers pour la détection des personnes dans des images et séquences d'images IR thermique est celui de Nanda and Davis, (2002), qui ont proposé de représenter les ROIs (extraites par un simple seuillage) en utilisant des modèles probabilistes, puis ils ont utilisé un classifieur Bayésien pour prédire la classe (humain ou non-humain) de ces ROIs. Davis and Keck, (2005) ont

proposé une méthode comportant deux étapes principales. Dan la première étape, une procédure de pré-sélection rapide basée sur des cartes de contours saillants (Contour Saliency Maps, CSM) est utilisée pour localiser les ROIs. Dans la deuxième étape, l'algorithme d'apprentissage AdaBoost est utilisé pour construire un ensemble de classifieurs capable de détecter des personnes. Suard et al., (2006) ont proposé d'utiliser le descripteur HOG avec un SVM comme classifieur pour la détection des personnes dans des images thermiques stéréo. Sun et al., (2011) ont proposé un descripteur appelé Motif Binaire Pyramidal (Pyramid Binarry Pattern, PBP) pour décrire la caractéristique de symétrie des corps humains dans des images thermiques. Ce descripteur est une extension du descripteur LBP en utilisant un agencement spatial des cellules de texture, et il a été combiné efficacement avec le classifieur SVM. Inspirés par le descripteur CSS (Color Self Similarity) introduit par Walk et al., (2010) pour la caractérisation de la couleur dans le spectre visible, Miron et al., (2012) ont proposé le descripteur ISS, qui est basé sur l'auto-similarité relative de l'intensité au sein des régions appartenant à une personne dans les images thermiques (à titre d'exemple, les pixels appartenant à la tête des personnes ont des valeurs d'intensité très similaires, correspondant ainsi à un degré de similarité relativement élevé). Zhao et al., (2015) ont proposé d'utiliser les CSMs pour extraire les ROIs susceptibles de contenir des personnes, puis ils ont appliqué le descripteur SDH sur ces régions afin de décrire leur forme. Une méthode de classification basée sur la représentation éparse modifiée (MSRC) est enfin utilisée par les auteurs pour réaliser la détection**.** Qi et al., (2016) ont aussi proposé une approche basée sur la représentation éparse pour la détection de personnes à partir des images thermiques. Dans cette approche, les auteurs ont tout d'abord adopté le descripteur HSC pour représenter les caractéristiques des régions candidates sélectionnées à l'aide de la méthode de la fenêtre glissante. Ensuite, les caractéristiques extraites sont introduites dans une méthode de classification basée sur la représentation éparse afin de déterminer la présence d'une personne. Plus récemment, les réseaux CNN ont été largement adoptés pour la détection de personnes dans des séquences d'images thermiques en raison de leurs performances élevées. John et al., (2015), par exemple, ont proposé d'appliquer l'algorithme des C-moyens flous (fuzzy C-means) sur les images d'entrée afin d'extraire les ROIs, puis ils ont utilisé les caractéristiques de posture humaine et l'ellipse d'inertie afin de réduire leur nombre. Les ROIs sélectionnées sont ensuite

redimensionnées à une taille fixe, puis elles sont transmises à un algorithme CNN à 8 couches pour effectuer la classification en humain ou non-humain. Herrmann et al., (2016) ont utilisé la méthode de détection des régions extrêmes à stabilité maximale (Maximally Stable Extremal Regions, MSER) pour extraire les régions de haute température (ROIs) à partir des images thermiques de faible résolution, puis ils ont appliqué un algorithme CNN sur ces régions afin de détecter la présence d'une personne. Heo et al., (2018) ont proposé d'utiliser l'algorithme ABMS (Adaptive Boolean-Map-based Saliency) pour mettre en évidence les régions des personnes (ROIs) à partir de l'arrière-plan de la scène, puis ils ont utilisé un détecteur nommé YOLOv2 (You Only Look Once version 2) (Redmon and Farhadi, 2017)**,** qui est basé sur les réseaux CNNs, pour permettre la détection des personnes en fonction du type de la saison. Les méthodes basées sur les CNNs sont capables de détecter des personnes avec précision. Cependant, elles sont souvent laborieuses, très coûteuses en termes de temps de calcul, et nécessitent de grands ensembles de données étiquetées pour entraîner les réseaux (Haq et al., 2020). Cela limite fortement leur applicabilité dans des scénarios réels.

2.3. Approches proposées

Dans cette section, nous proposons deux approches différentes pour la détection automatique de personnes en mouvement dans des séquences d'images IR. La première est basée sur le calcul d'une fonction de similarité combinée qui utilise des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement. La deuxième est basée sur la détection conjointe des deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre majuscule de l'alphabet grec Omega Ω), et les deux jambes. Cependant, avant de présenter les détails de ces deux approches proposées, nous commençons tout d'abord dans la sous-section suivante par présenter les différentes opérations de prétraitement appliquées sur les séquences d'images IR afin de les préparer aux étapes ultérieures dans le processus de détection de personnes.

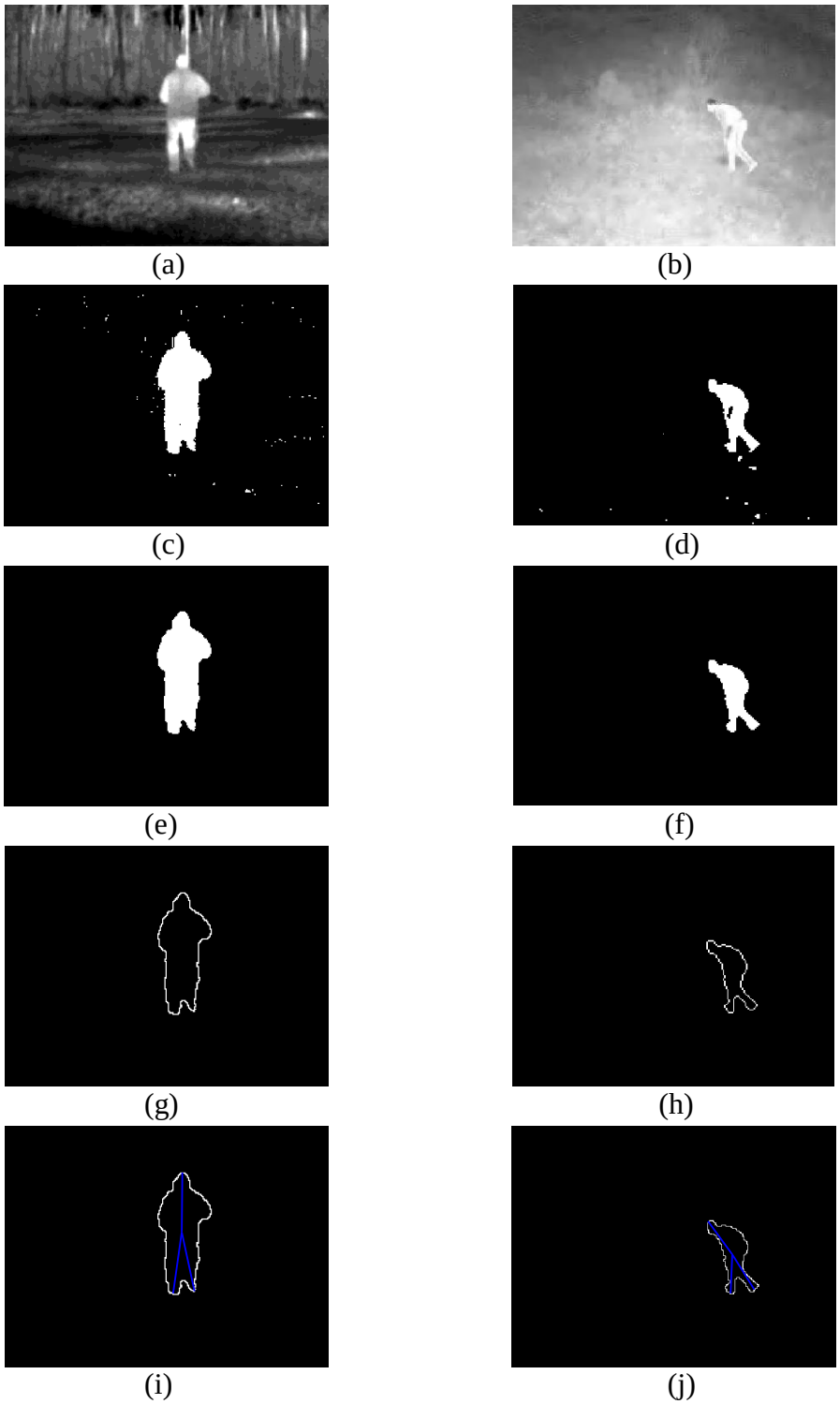


Figure 2.2: Opérations de prétraitement (colonne de gauche : image IR thermique, colonne de droite : image proche IR). (a-b) Trames originales. (c-d) Résultat d'extraction d'objets en mouvement en utilisant la méthode GMM adaptative. (e-f) Résultat de rehaussement. (g-h) Résultat d'extraction du contour. (i-j) Résultat d'extraction du squelette-étoile.

2.3.1. Prétraitements

Les opérations de prétraitement comprennent : 1) l'extraction des objets en mouvement (ROIs), 2) rehaussement du masque d'avant-plan, 3) extraction du contour, et 4) extraction du squelette-étoile. Les détails sur chacune de ces opérations sont décrits ci-dessous.

2.3.1.1. Extraction des objets en mouvement

La première opération de prétraitement consiste à extraire les objets en mouvement à partir des séquences d'images IR. Parmi les techniques les plus populaires pour réaliser cette tâche, nous pouvons citer le flux optique, la dérivée temporelle et la méthode de soustraction d'arrière-plan. Dans ce travail, comme nous travaillons avec des séquences d'images IR acquises par des caméras de surveillance statiques, nous avons adopté la méthode de soustraction d'arrière-plan adaptative basée sur le modèle de mélange de gaussiennes (GMM) (Goyal and Singhai, 2018), qui fonctionne de manière robuste dans des environnements extérieurs réels contenant des situations difficiles telles que des changements d'illumination, des encombrements (clutters), et des changements à long terme dans la scène. Le principe de cette méthode est de modéliser l'historique récent $\{I_1, \ldots, I_k\}$ de l'intensité de chaque pixel de l'image par un mélange de K distributions gaussiennes. Ces distributions sont ensuite évaluées en utilisant des règles heuristiques simples afin de déterminer lesquelles d'entre elles représentent l'arrière-plan. La probabilité d'observer une valeur d'un pixel I_k à l'instant k est estimée par la l'équation suivante :

$$P(I_k) = \sum_{i=1}^{K} \mathcal{W}_{i,k} \mathcal{N}(I_k;\, \mu_{i,k}, \Sigma_{i,k}) \tag{2.1}$$

où K (typiquement choisi de 3 à 5) est le nombre de distributions gaussiennes, $\mathcal{W}_{i,k}$ est l'estimation à l'instant k du poids assigné à la i-ième gaussienne dans le mélange, $\mu_{i,k}$ est sa valeur moyenne, $\Sigma_{i,k}$ est sa matrice de covariance supposée égale à $\sigma_{i,k}^2\mathbf{I}$ (**I** est la matrice identité), et $\mathcal{N}$ est la Fonction Densité de Probabilité (PDF) gaussienne définie par:

$$\mathcal{N}(I_k;\ \mu,\Sigma) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(I_k-\mu)^{\mathrm{T}}\Sigma^{-1}(I_k-\mu)} \tag{2.2}$$

Afin de déterminer le type (c.-à-d., arrière-plan ou avant-plan) du pixel courant, les distributions gaussiennes du modèle de mélange sont tout d'abord rangées par ordre décroissant en fonction de la valeur de la fitness ($F_{i,k}$) définie dans l'équation suivante :

$$F_{i,k} = \frac{\mathcal{W}_{i,k}}{\left\|\Sigma_{i,k}\right\|} \tag{2.3}$$

Ensuite, les B premières distributions sont sélectionnées et utilisées comme un modèle de l'arrière-plan, avec B est défini comme suit :

$$B = \arg\min_{b}\left(\sum_{i=1}^{b} \mathcal{W}_{i,k} > \tau\right) \tag{2.4}$$

où $\tau > 0$ est un seuil qui représente la fraction minimale du modèle d'arrière-plan. Si une petite valeur est choisie pour τ, le modèle d'arrière-plan est généralement uni-modal.

Les pixels d'avant-plan représentant les objets en mouvement sont obtenus en utilisant la formule de l'équation 2.5, où κ est une valeur constante, souvent fixée à 2.5.

$$\sqrt{\left(I_k - \mu_{i,k}\right)^{\mathrm{T}} \Sigma_{i,k}^{-1} \left(I_k - \mu_{i,k}\right)} < \kappa\sigma_{i,k} \tag{2.5}$$

Une fois les objets d'avant-plan sont extraits, les poids $\mathcal{W}_{i,k}$ ainsi que les paramètres statistiques $\{\mu_{i,k}, \sigma_{i,k}^2\}$ du modèle de mélange de gaussiennes sont finalement mis à jour afin de s'adapter aux changements dynamiques de la scène. Les équations pour effectuer cette mise à jour peuvent être trouvées dans (Goyal and Singhai, 2018).

Des exemples montrant les résultats d'extraction d'objets en mouvement à partir d'une image thermique (Figure 2.2(a)) et d'une image proche IR (Figure 2.2(b)) en utilisant la méthode de soustraction d'arrière-plan basée sur le GMM sont illustrés dans les Figures 2.2(c) et 2.2(d).

2.3.1.2. Rehaussement du masque d'avant-plan

Très souvent, et plus particulièrement dans les environnements réels extérieurs de nuit, le masque d'avant-plan binaire obtenu après avoir effectué la soustraction d'arrière-plan n'est pas parfait et contient une certaine quantité de pixels erronément détectés comme en mouvement. De plus, les régions appartenant aux objets d'avant-plan sont généralement corrompues par quelques trous et certaines parties défaillantes, qui sont principalement causés par des pixels d'objets en mouvement possédant des valeurs d'intensité similaires à celles de l'arrière-plan. Ainsi, afin de réduire ces erreurs, nous avons appliqué sur le masque d'avant-plan une cascade d'opérateurs morphologiques d'ouverture et de fermeture suivie d'un filtre médian. L'opération d'ouverture a pour objectif de supprimer les pixels bruyants isolés du masque d'avant-plan, alors que l'opération de fermeture a pour objectif de remplir les petits trous à l'intérieur des silhouettes d'avant-plan et de fusionner les régions divisées après le processus de soustraction d'arrière-plan. Le filtre médian est ajouté pour supprimer le bruit impulsif restant après l'application du filtrage morphologique, et ce, sans détériorer les silhouettes des objets en mouvement.

Cependant, dans la pratique, malgré que l'application des opérateurs morphologiques améliore considérablement la qualité des silhouettes d'objets en mouvement, quelques petits blobs non désirables peuvent encore exister dans le masque d'avant-plan. Ainsi, afin de réduire au maximum ces blobs, nous avons appliqué l'algorithme d'étiquetage en composantes connexes (Connected Component Labelling, CCL) sur le masque d'avant-plan, puis, toute composante connexe ayant une aire inférieure à un seuil donné (représentant la taille minimale possible pour un être humain dans la scène) est considérée comme étant un *outlier* et elle est supprimée du masque d'avant-plan.

Les résultats d'opération de rehaussement appliquée sur les masques d'avant-plan des Figures 2.2(c) et 2.2(d) sont montrés dans les Figures 2.2(e) et 2.2(f).

2.3.1.3. Extraction du contour

Après l'extraction des objets en mouvement, la prochaine opération de prétraitement consiste à extraire les contours des silhouettes de ces objets, et qui seront utilisés ultérieurement dans le processus de détection de personnes et de reconnaissance de leur posture. Pour ce faire, nous avons utilisé l'algorithme de

traçage de contours de voisinage de Moore modifié par le critère d'arrêt de Jacob (Pradhan et al., 2010). Cet algorithme est décrit comme suit.

Rappelons que la notion de voisinage de Moore d'un pixel d'une image comme est l'ensemble des huit pixels qui partagent un vertex ou un bord avec ce pixel. Soit une image binaire contenant une composante connexe (l'objet en mouvement) avec des pixels blancs sur un arrière-plan noir. L'algorithme de traçage de contour de voisinage de Moore commence par balayer l'image de gauche à droite et de haut en bas jusqu'à ce qu'un pixel blanc soit rencontré. Ce pixel est défini comme le pixel de départ pour l'algorithme de traçage de contour. Ensuite, le voisinage de Moore de ce pixel est examiné dans le sens des aiguilles d'une montre jusqu'à ce que le prochain pixel blanc soit trouvé. Ce processus recommence à chaque fois qu'un nouveau pixel blanc est trouvé et l'algorithme s'arrête lorsque le pixel de départ est visité pour une deuxième fois (c'est le critère d'arrêt de Jacob). Finalement, l'ensemble des pixels blancs visités au cours du processus de balayage constitue le contour de l'objet présent dans l'image.

Les résultats de l'application de l'algorithme de traçage de contour décrit ci-dessus sur les masques d'avant-plan des Figures 2.2(e) et 2.2(f) sont montrés aux Figures 2.2(g) et 2.2(h) précédentes.

2.3.1.4. Extraction du squelette-étoile

Le concept de base du squelette-étoile (Star-skeleton, en Anglais) (Afsar et al., 2017) est de localiser initialement les points extrêmes du contour de la silhouette, et de les connecter ensuite au centre de gravité de cette silhouette. Parmi les avantages les plus intéressants de ce descripteur sont sa simplicité et son coût de calcul très faible. Cependant, sa qualité de représentation est très dépendante de la qualité du contour de la silhouette qui, dans beaucoup de situations, n'est pas parfait et contient une certaine quantité de bruits. Ainsi, un filtrage du contour de la silhouette est nécessaire afin de garantir une extraction efficace du squelette-étoile. Dans ce travail, pour accomplir cette tâche, nous avons employé le filtrage par la transformée de Fourier discrète (TFD). La procédure est décrite ci-dessous.

Si nous notons $z_i = x_i + jy_i$, les coordonnées complexes du contour de la silhouette à filtrer, avec $\{x_i, y_i\}_{i=0}^{N_c-1}$ sont les coordonnées le long des axes des x et y du i-éme point de contour, et N_c est le nombre total de points dans le contour. La transformée de Fourier discrète de z_i est calculée en utilisant l'expression suivante :

$$C_k = \frac{1}{N_c}\sum_{i=0}^{N_c-1} z_i e^{\left(-j\,\frac{2\pi\, i\, k}{N_c}\right)}, \quad k = 0, \ldots, N_c - 1 \tag{2.6}$$

où, les coefficients $\{C_k\}_{k=0}^{N_c-1}$ sont les descripteurs de Fourier (DFs) (Larsson and Felsberg, 2011) du contour. Ainsi, si nous retenons uniquement les M (un entier appartenant à $[1, N_c/2]$) premiers coefficients d'ordre inferieur et les M derniers coefficients d'ordre supérieur, et que nous mettons tous les autres coefficients restants à zéro, la forme approximative (ou filtrée) du contour original peut être obtenue par la transformée de Fourier inverse qui est donnée par l'expression 2.7.

$$\tilde{z}_i = \frac{1}{N_c}\sum_{k=0}^{N_c-1} \tilde{C}_k e^{\left(j\,\frac{2\pi\, i\, k}{N_c}\right)}, \quad i = 0, \ldots, N_c - 1 \tag{2.7}$$

où, $\{\tilde{C}_k\}_{k=0}^{N_c-1}$ est le vecteur contenant les M premiers coefficients d'ordre inférieur et les M derniers coefficients d'ordre supérieur, avec le reste des coefficients mis à zéro. En variant la valeur du paramètre M, nous pouvons ajuster la quantité de bruit et les petits détails supprimés du contour de la silhouette.

Un exemple de filtrage d'un contour humain en utilisant la transformée de Fourier discrète avec différentes valeurs du paramètre M est montré à la Figure 2.3. À partir de cette figure, nous pouvons observer que, plus la valeur du paramètre M est grande, plus les détails du contour sont retenus. En d'autres termes, plus la valeur du paramètre M est petite, plus les détails sont filtrés. Dans nos expériences, pour déterminer la valeur correcte du paramètre M, nous avons effectué plusieurs tests avec plusieurs valeurs, et celle correspondant à la meilleure performance de l'algorithme de détection de personnes est retenue comme valeur "optimale".

Figure 2.3: Exemple de filtrage du contour d'une silhouette humaine en utilisant la transformée de Fourier discrète avec différentes valeurs du paramètre M.

Une fois la tâche de filtrage du contour accomplie, l'extraction du squelette-étoile peut être effectuée. Supposons tout d'abord que nous parcourons le contour filtré dans le sens des aiguilles d'une montre. Ensuite, commençons par le point de contour le plus à gauche, puis calculons la distance Euclidienne, définie par l'expression 2.8, de chaque point du contour au centre de gravité de la silhouette.

$$d_i = \sqrt{(\tilde{x}_i - x_G)^2 + (\tilde{y}_i - y_G)^2}, \qquad i = 0,1, \dots, N_c - 1 \tag{2.8}$$

Dans l'équation ci-dessus, $\{\tilde{x}_i, \tilde{y}_i\}_{i=0}^{N_c-1}$ dénotent les coordonnées des points présents sur le contour filtré, et (x_G, y_G) dénotent les coordonnées du centre de gravité de la silhouette.

Après avoir calculé la fonction de distance Euclidienne d_i, l'étape suivante consiste à trouver ses maxima locaux, qui correspondent aux points saillants du contour de la silhouette. Cependant, dans notre contexte d'application (détection de personnes), et afin d'éviter d'extraire plusieurs points saillants du contour, nous retenons uniquement les points suffisamment éloignés du centre de gravité de la silhouette. Ainsi, les étapes pour obtenir ces points saillants sont les suivantes :

Etape 1 : Normaliser la fonction de distance $\{d_i\}_{i=0}^{N_c-1}$ par rapport à sa valeur maximale, d_{max}, pour obtenir la fonction de distance normalisée, $\hat{d}_i$:

$$\hat{d}_i = \frac{d_i}{d_{max}}, \qquad i = 0,1, \dots, N_c - 1 \tag{2.9}$$

Cela résulte en $0 \leq \hat{d}_i \leq 1$.

Le but de cette normalisation est d'atteindre l'invariance à l'échelle.

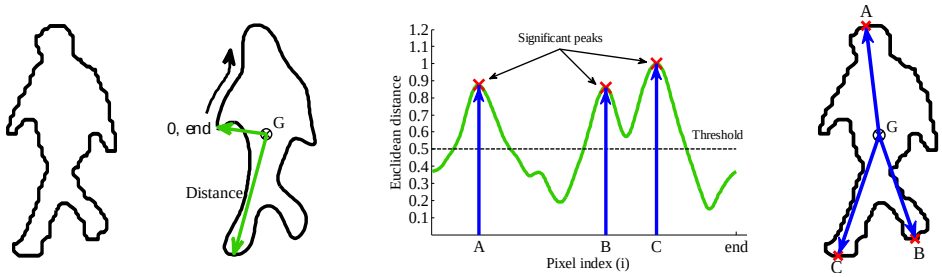
Étape 2 : Déterminer les maxima locaux de $\hat{d}_i$ en localisant les passages par zéro de la fonction de différence suivante :

$$\delta_i = \hat{d}_i - \hat{d}_{i-1}, \qquad i = 1, \dots, N_c - 1 \tag{2.10}$$

Étape 3 : Conserver comme points saillants les points de contour correspondant aux maxima locaux de $\hat{d}_i$ ayant une amplitude supérieure à un seuil prédéfini (0.5 dans nos expériences).

Après avoir obtenu les points saillants du contour, le squelette-étoile peut être finalement construit en reliant ces points au centre de gravité de la silhouette. Une

illustration de la procédure globale d'extraction du squelette-étoile pour un exemple de contour humain est montrée dans la Figure 2.4.



(a) (b) (c) (d)

Figure 2.4 : Procédure d'extraction du squelette-étoile pour un exemple de contour humain. (a) Contour original. (b) Contour filtré en utilisant la transformée de Fourier (M = 12). (c) Fonction de distance Euclidienne normalisée (les maxima locaux les plus significatifs sont indiqués par les croix rouges). (d) Le squelette-étoile obtenu.

Les résultats d'extraction du squelette-étoile pour les objets en mouvement dans les Figure 2.2(g-h) sont montrés dans les Figures 2.2(i-j).

2.3.2. Première approche proposée

La première approche de détection de personnes que nous proposons est basée sur le calcul d'une fonction de similarité globale pour chaque objet en mouvement extrait à partir de l'arrière-plan de la scène observée. Cette fonction de similarité globale est une combinaison de plusieurs sous-fonctions de similarité individuelles, qui sont : 1) la similarité basée sur le squelette-étoile, 2) la similarité basée sur le rapport de forme, 3) la similarité basée sur le rapport arrière-plan/avant-plan, 4) la similarité basée sur la distance spatiale, et 5) la similarité basée sur le rapport de chevauchement. Chacune de ces sous-fonctions de similarité est décrite dans les sous-sections suivantes.

2.3.2.1. Similarité basée sur le squelette-étoile

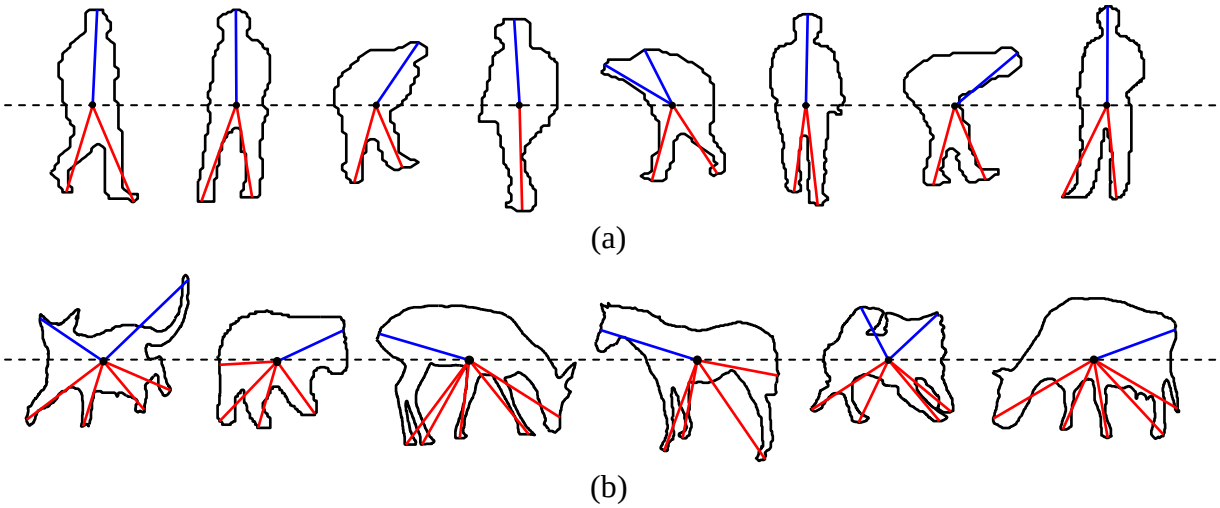
Le squelette-étoile est une caractéristique essentielle pour la représentation du corps humain dans le plan image. Comparé à celui d'autres objets en mouvement tels que les animaux, le squelette-étoile du corps humain est souvent caractérisé par la présence d'un ou de deux segments de ligne dans sa partie

inférieure (Figure 2.5), sachant que la grande majorité des animaux se déplacent sur leurs quatre pattes. Ainsi, afin de distinguer un être humain à partir d'autres objets en mouvement, dans cette première approche que nous proposons, nous définissons la similarité S_{Skel} basée sur le squelette-étoile comme suit :

$$S_{Skel} = \begin{cases} 1 & \text{si } 1 \text{ ou } 2 \text{ segments de lignes sont détéctés à l'interieur de} \\ & \text{l'intervalle } [-\theta_{legs}, +\theta_{legs}] \\ 0 & \text{autrement} \end{cases} \quad (2.11)$$

Cela signifie que si un segment de ligne (correspondant à des jambes humaines fermées) ou deux segments de ligne (correspondant à des jambes humaines séparées) sont détectés dans la partie inférieure du squelette-étoile, spécifiée par l'intervalle d'angle $[-\theta_{legs}, +\theta_{legs}]$, illustré dans la Figure 2.6. La valeur de la similarité S_{Skel} est mise à 1. Si non, sa valeur est mise à 0.

À partir de la définition dans l'équation 2.11, nous pouvons constater que le choix des valeurs pour l'intervalle $[-\theta_{legs}, +\theta_{legs}]$ aura une grande influence sur les performances globales de la méthode proposée. Dans nos expériences, les valeurs "optimales" pour cet intervalle sont fixées après plusieurs tests sur un ensemble de données contenant des silhouettes humaines et non humaines.



(a)

(b)

Figure 2.5 : Le squelette-étoile pour (a) des silhouettes humaines et (b) des silhouettes non humaines (animaux). La ligne horizontale discontinue représente la limite entre la partie supérieure et la partie inferieure du squelette-étoile.

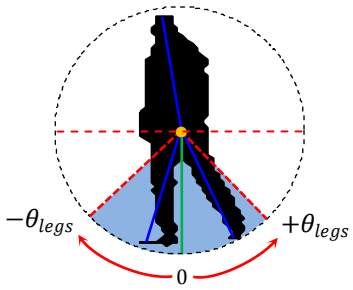


Figure 2.6: L'intervalle $[-\theta_{legs}, +\theta_{legs}]$ spécifiant la partie inférieure du squelette-étoile.

2.3.2.2. Similarité basée sur le rapport de forme

Le rapport de forme (AR), encore appelé rapport d'aspect (Aspect Ratio), est défini comme une mesure de l'élongation et de la minceur d'un objet dans l'image. Comme le montre l'équation 2.12, il est calculé comme étant le rapport entre la hauteur H_{MBB} et la largeur L_{MBB} de la boîte englobante minimale (Minimum Bounding Box, MBB) qui entoure tous les pixels constituant l'objet d'intérêt dans l'image.

$$AR = \frac{H_{MBB}}{L_{MBB}} \tag{2.12}$$

La valeur de ce rapport est généralement supérieure à 1 pour un être humain en mouvement, et elle est inférieure à 1 pour d'autres objets non humains en mouvement. Ainsi, dans ce travail, comme nous nous intéressons uniquement à la détection de personnes en mouvement, nous définissons la mesure de similarité, S_{AR}, basée sur le rapport de forme, comme suit :

$$S_{AR} = \begin{cases} \frac{AR}{AR_{Max}} & si \quad AR_{Min} \leq AR \leq AR_{Max} \\ 0 & autrement \end{cases} \tag{2.13}$$

où AR_{Min} et AR_{Max} sont les valeurs minimales et maximales possibles du rapport AR pour un être humain dans la scène surveillée. Dans nos expériences, ces valeurs sont fixées, respectivement, à 1 et 4. L'équation 2.13 montre que plus la valeur de S_{AR} est élevée, plus il est probable que l'objet en mouvement détecté soit un être humain.

2.3.2.3. Similarité basée sur le rapport arrière-plan/avant-plan

Le rapport arrière-plan/avant-plan, η, est calculé comme le rapport entre la valeur moyenne de l'intensité de la région locale d'arrière-plan qui entoure l'objet en mouvement détecté et la valeur moyenne de l'intensité de sa région d'avant-plan. Dans ce travail, comme illustré sur la Figure 2.7, la région locale d'arrière-plan est choisie comme étant la boîte englobante minimale qui entoure l'objet en mouvement multipliée par un facteur d'échelle $\lambda > 1$. Toutefois, afin de différencier plus clairement un être humain d'un autre objet non humain en mouvement (tel qu'un animal, un véhicule, etc.), la valeur de l'intensité moyenne de l'objet en mouvement détecté est pondérée par la valeur de son rapport de forme (AR), définie précédemment dans l'équation 2.12. Ainsi, la formule pour le calcul du rapport arrière-plan/avant-plan, η, peut être exprimée comme suit :

$$\eta = \frac{\left(\frac{1}{N_{Back}} \sum_{i \in Back} I(i)\right)}{\left(\frac{1}{N_{Fore}} \sum_{j \in Fore} I(j)\right)} \times \frac{1}{AR} \tag{2.14}$$

où $I(i)$ est la valeur d'intensité au i-ème pixel de l'image IR, N_{Fore} et N_{Back} sont, respectivement, le nombre de pixels contenus dans la région d'avant-plan et la région locale d'arrière-plan de l'objet en mouvement.

Ainsi, à partir du rapport arrière-plan/avant-plan, η, défini dans l'équation 2.14, nous définissons la similarité S_{Back_Fore} basée sur ce rapport. Elle est calculée par l'expression 2.15.

$$S_{Back\_Fore} = \begin{cases} 1 - \eta & \text{si} \quad 0 \leq \eta \leq 1 \\ 0 & autrement \end{cases} \tag{2.15}$$

Comme le corps humain dans les images IR se caractérise par des valeurs d'intensité élevées (Figure 2.7), la valeur de la similarité S_{Back_Fore} sera très proche de 1 pour un être humain, et inversement, elle sera très proche de 0 pour d'autres objets non humains en mouvement.

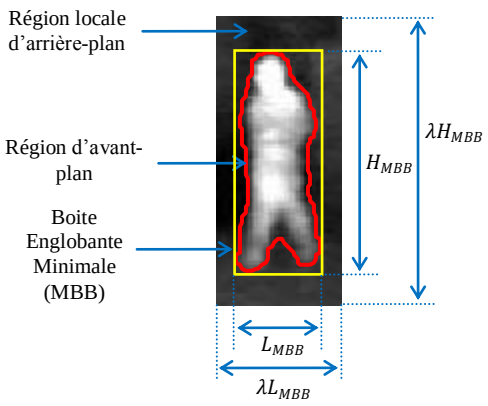


Figure 2.7: Exemple d'un objet d'avant-plan, sa boite englobante minimale, et sa région locale d'arrière-plan.

2.3.2.4. Similarité basée sur la distance spatiale

Cette similarité, définie dans l'équation 2.16, est calculée sur la base de la distance linéaire entre les centres de gravité (Centers-of-Gravity, CoGs) d'un objet en mouvement détecté à deux instants consécutifs k et $k-1$.

$$S_{Spatial} = 1 - \frac{d(\mathrm{CoG}(O_k),\ \mathrm{CoG}(O_{k-1}))}{R_s} \tag{2.16}$$

où $d(\mathrm{CoG}(O_k),\ \mathrm{CoG}(O_{k-1}))$ est la distance Euclidienne entre les CoGs d'un objet en mouvement détecté aux instants k et $k-1$, et R_s définit le rayon de voisinage de recherche autour de la position de l'objet O_{k-1} détecté à l'instant $k-1$.
La valeur de ce rayon utilisée dans nos expériences est choisie comme étant la distance maximale du CoG de l'objet O_{k-1} aux points extrêmes de son squelette-étoile (Figure 2.6).
L'équation 2.16 montre que, plus les centres $\mathrm{CoG}(O_k)$ et $\mathrm{CoG}(O_{k-1})$ sont proches l'un de l'autre, plus la valeur de la similarité $S_{Spatial}$ est proche de 1, et par conséquent, plus il est probable que les deux objets en mouvement O_{k-1} et O_k sont les mêmes dans des trames différentes.

2.3.2.5. Similarité basée sur le rapport de chevauchement

Le rapport de chevauchement $S_{Overlap}$ est une mesure de similarité qui quantifie le degré de chevauchement entre les boîtes englobantes minimales d'un

objet en mouvement détecté à deux instants consécutifs $k-1$ et k. Sa valeur est calculée par l'expression suivante :

$$S_{Overlap} = \frac{Area(MBB(O_k) \cap MBB(O_{k-1}))}{Area(MBB(O_k) \cup MBB(O_{k-1}))} \quad (2.17)$$

où $MBB(O_k)$ et $MBB(O_{k-1})$ sont les boîtes englobantes minimales d'un objet en mouvement détecté aux instants k et $k-1$, $\cup$et $\cap$ sont, respectivement, l'opérateur d'union et d'intersection, et la fonction $Area\,(\cdot)$ signifie calculer l'aire (en pixels) d'une région.

La Figure 2.8 illustre la zone de chevauchement pour un exemple d'objet en mouvement détecté à deux trames consécutives. L'équation 2.17 montre que plus le degré de chevauchement entre $MBB(O_k)$ et $MBB(O_{k-1})$ est élevé, plus la valeur de la similarité $S_{Overlap}$ est proche de 1, et par conséquent, plus il est probable que les deux objets en mouvement O_{k-1} et O_k sont les mêmes dans des trames différentes.

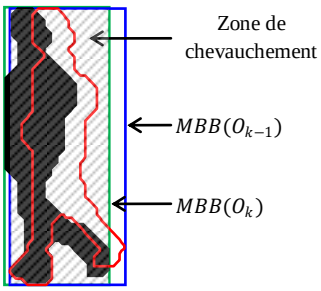


Figure 2.8: Zone de chevauchement (zone hachurée) entre les boites englobantes minimales d'un objet en mouvement détecté à deux trames consécutives.

2.3.2.6. Similarité globale

La fonction de similarité globale, S_{Global} , utilisée pour la détection de personnes peut être finalement calculée à l'aide de la formule donnée dans l'équation 2.18, en combinant les différentes sous-fonctions de similarité individuelles définies précédemment.

$$S_{Global} = S_{Skel}\left(S_{AR} + S_{Back\_Fore} + S_{Spatial} + S_{Overlap}\right) \quad (2.18)$$

Ainsi, afin de distinguer un être humain à partir d'autres objets non humains en mouvement, nous proposons d'appliquer un seuil, T_{Sim} , comme le montre

l'expression 2.19, sur la fonction de similarité globale S_{Global} . Tout objet en mouvement qui satisfait cette condition est ensuite considéré comme un être humain. Autrement, il est considéré comme un objet non humain en mouvement. La valeur "optimale" du seuil T_{Sim} est déterminée empiriquement en cherchant à maximiser la performance de détection.

$$\begin{cases} "Humain" & \text{si} \quad S_{Global} \geq T_{Sim} \\ "Non-humain" & \quad autrement \end{cases} \tag{2.19}$$

Cependant, dans le cas pratique, notamment dans des environnements extérieurs de surveillance, nous avons observé que l'utilisation seule de la condition de l'équation 2.19 pour détecter des personnes en mouvement peut facilement être affectée par de petits mouvements (clutters), tels que le balancement des branches d'arbres, les fluctuations d'eau, le vol d'oiseaux, etc., qui se produisent dans l'arrière-plan. Ceci peut augmenter considérablement le taux de fausses détections. Ainsi, pour surmonter ce problème, dans cette première approche, seuls les objets en mouvement satisfaisant la condition de l'équation 2.19 sur un intervalle de temps prédéfini contenant un certain nombre de trames consécutives, que nous notons n_f, sont considérés comme des êtres humains. Autrement, tout autre objet qui ne satisfait pas cette condition est considéré comme une fausse alarme et il est exclu de l'étape de suivi.

2.3.3. Deuxième approche proposée

La deuxième approche que nous proposons pour la détection de personnes est basée sur la détection conjointe des deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre majuscule de l'alphabet grec Omega Ω), et les deux jambes. Cette approche est composée principalement de trois étapes. La première consiste en l'extraction de la meilleure ellipse ajustée. La deuxième consiste à la détection de la partie tête-épaules. Quant à la troisième, elle consiste à la détection des jambes.
Les détails sur chacune de ces trois étapes sont décrits dans les sous-sections suivantes.

2.3.3.1. Extraction de la meilleure ellipse ajustée

La meilleure ellipse ajustée (best-fitting ellipse) pour une silhouette d'un objet est définie comme étant l'ellipse équivalente ayant les mêmes moments centraux d'ordre 2 que cette silhouette. Notons que, les moments centraux d'ordre (p, q) pour une image $I(x, y)$ sont définis comme suit :

$$\mu_{p,q} = \sum_{x}\sum_{y}(x - x_G)^p (y - y_G)^q I(x, y)\,, \qquad p, q = 0,1,2, \ldots \tag{2.20}$$

où (x, y) dénotent les coordonnées des pixels contenus dans la silhouette, et (x_G, y_G) dénotent les coordonnées du centre de gravité de cette silhouette. Ainsi, sur la base des moments centraux d'ordre 0, 1 et 2, les longueurs (l, w) de l'axe semi-majeur et l'axe semi-mineur de la meilleure ellipse ajustée de la silhouette peuvent être calculées comme suit (Rocha et al., 2004) :

$$l = 2\sqrt{\frac{I_1}{\mu_{0,0}}} \tag{2.21.a}$$

$$w = 2\sqrt{\frac{I_2}{\mu_{0,0}}} \tag{2.21.b}$$

où, I_1 et I_2 sont définis par :

$$I_1 = \frac{(\mu_{2,0} + \mu_{0,2}) + \sqrt{(\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2}}{2} \tag{2.22.a}$$

$$I_2 = \frac{(\mu_{2,0} + \mu_{0,2}) - \sqrt{(\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2}}{2} \tag{2.22.b}$$

L'angle θ, $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, que fait l'axe semi-majeur de la meilleure ellipse ajustée avec l'axe des abscisses (x) peut aussi être calculé en utilisant les moments centraux d'ordre 2, comme montré dans l'équation 2.23 :

$$\theta = \begin{cases} \theta_0 & \text{si} \quad \mu_{2,0} > \mu_{0,2} \\ \theta_0 + \text{sign}(\mu_{1,1})\dfrac{\pi}{2} & \text{si} \quad \mu_{2,0} < \mu_{0,2} \end{cases} \tag{2.23}$$

où, θ_0 est défini par:

$$\theta_0 = \frac{1}{2}\tan^{-1}\left[\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}\right] \quad (2.24)$$

Les paramètres (l, w, θ) de la meilleure ellipse ajustée donnent ainsi une bonne estimation de la taille et de l'orientation de la silhouette de l'objet dans le plan image.

2.3.3.2. Détection de la partie tête-épaules

La détection de la partie tête-épaules humaine comprend trois étapes principales. La première consiste à extraire les contours tête-épaules candidats de la silhouette de l'objet en mouvement, et ce en se basant sur sa meilleure ellipse ajustée et son squelette-étoile. La deuxième étape consiste à passer ces contours candidats à travers une phase d'extraction des caractéristiques afin d'extraire leurs caractéristiques qui décrivent leur forme. En dernière étape, nous introduisons ces caractéristiques dans un classifieur SVM entraîné pour détecter la présence de la partie tête-épaules humaine au sein de la silhouette de l'objet en mouvement. Ces différentes étapes sont décrites dans les sous-sections suivantes.

2.3.3.2.1. Extraction des contours tête-épaules candidats

La deuxième approche de détection des personnes que nous proposons se base sur deux observations principales. La première est que, la forme tête-épaules humaine est la partie la plus invariante par rapport aux changements d'angle de vue et aux déformations du corps humain. La deuxième est que, cette partie est généralement la partie la moins affectée par les occultations par rapport à d'autres parties du corps humain. Lorsqu'un être humain est en mouvement constant, certaines parties de son corps, telles que la tête, les épaules et la région du torse restent relativement stables, tandis que d'autres parties telles que les bras et les jambes tendent à subir des mouvements importants dans différentes directions. Typiquement, de tels mouvements entraînent de grandes variations de l'apparence du corps humain, particulièrement autour du torse et de la partie inférieure. A titre d'exemple, comme l'illustre bien la Figure 2.9, une silhouette humaine avec des bras complètement visibles apparaît différente d'une silhouette avec des bras partiellement ou complètement occultés par le torse. De plus, dans des circonstances normales, la partie tête-épaule humaine n'interagit pas physiquement

avec l'environnement, contrairement à d'autres parties du corps, notamment les bras, et moins fréquemment les jambes, qui peuvent interagir entre elles et/ou avec le milieu environnant. L'ensemble de toutes ces caractéristiques fait de la partie tête-épaules une partie très discriminante pour distinguer un être humain à partir d'autres objets en mouvement dans un environnement extérieur de nuit.

(a) (b)

Figure 2.9 : Silhouette humaine avec (a) des bras partiellement occultés, et (b) des bras complètement visibles.

Le principal défi que nous avons à relever est de savoir comment détecter de manière robuste la partie tête-épaules humaine indépendamment des variations de la posture humaine, de l'échelle et de l'angle de vue de la caméra. Pour ce faire, nous proposons dans cette approche d'utiliser les deux caractéristiques décrites dans les sous-sections 2.3.1.4 et 2.3.3.1, à savoir le squelette-étoile et la meilleure ellipse ajustée. Pour illustrer notre procédure, considérons l'exemple du contour de silhouette présenté sur la Figure 2.10, où les segments de lignes bleues représentent les branches du squelette-étoile formé en reliant les points saillants du contour (notés par les lettres "A", "B" et "C") au centre de gravité de la silhouette (noté par la lettre "G"), l'ellipse en pointillé magenta représente la meilleure ellipse ajustée de la silhouette, et les lignes (M1, M2) et (m1, m2) sont, respectivement, ses axes majeur et mineur.

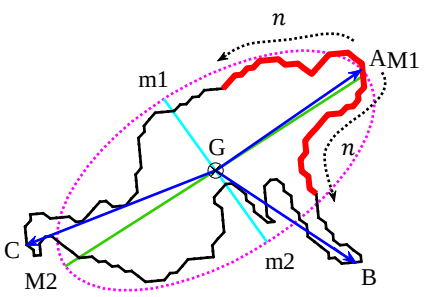


Figure 2.10 : Extraction des contours tête-épaules candidats.

Pour procéder à l'extraction des contours tête-épaules candidats, nous proposons de diviser initialement la silhouette détectée en une partie supérieure et une partie inférieure en utilisant l'axe mineur de la meilleure ellipse ajustée. Comme la partie supérieure du corps humain est souvent la partie la plus susceptible de contenir la forme tête-épaules par rapport la partie inférieure, nous proposons de commencer la recherche de la forme tête-épaules humaine de la partie supérieure à la partie inférieure de la silhouette de l'objet détecté.

L'extraction des contours tête-épaules candidats peut ainsi être effectuée en utilisant la procédure à deux étapes.

Dans la première étape, nous partons de l'axe semi-majeur supérieur et balayons la partie supérieure de la silhouette dans le sens des aiguilles d'une montre et dans le sens opposé. Ensuite, les points saillants du contour correspondant à la première branche du squelette-étoile trouvée dans chaque direction sont choisis comme des emplacements pour extraire les contours tête-épaules candidats. Si le nombre total de points saillants trouvés est égal à deux, alors les points sont triés dans l'ordre croissant, en fonction de l'angle entre leurs branches correspondantes du squelette-étoile et l'axe semi-majeur supérieur. Le point saillant du contour ayant l'angle minimal est sélectionné comme le premier emplacement pour extraire un contour tête-épaules candidat, tandis que le point restant est choisi comme le deuxième emplacement. Une fois que cette procédure est exécutée pour la partie supérieure de la silhouette, elle est répétée pour la partie inférieure afin de trouver deux autres emplacements pour extraire les contours tête-épaules candidats. A la fin de la procédure, tous les points saillants du contour obtenus à partir de la partie supérieure et la partie inférieure de la silhouette sont envoyés vers la deuxième étape pour extraire leurs contours tête-épaules candidats correspondants.

L'application de cette première étape sur le contour de la silhouette de la Figure 2.10 donne le résultat suivant. Deux points saillants du contour, à savoir, les points "A" et "B", qui correspondent aux branches (G, A) et (G, B) du squelette-étoile sont trouvés dans la partie supérieure de la silhouette pour extraire les contours tête-épaules candidats, alors qu'un seul point saillant, à savoir le point "C", qui correspond à la branche (G, C) du squelette-étoile, est trouvé dans la partie inférieure de la silhouette.

La deuxième étape implique l'extraction des contours tête-épaules candidats correspondant aux points saillants trouvés dans l'étape 1. Pour illustrer notre

procédure, considérons l'exemple du point saillant "A" dans le contour de la Figure 2.10, qui est trouvé dans l'étape 1 comme étant le premier point pour extraire le contour tête-épaules candidat. Ensuite, les n points de contour en partant du point "A" dans le sens des aiguilles d'une montre et dans le sens opposé le long du contour sont extraits comme un contour candidat pour détecter la partie tête-épaules humaine. La valeur du paramètre n dans nos expériences est calculée automatiquement, en fonction de la longueur totale L_c (en pixels) du contour de la silhouette par l'expression 2.25.

$$n = \frac{\alpha L_c}{2} \tag{2.25}$$

où α est le rapport des pixels tête-épaules humains par rapport à la longueur totale du contour du corps humain. Sa valeur est fixée à 0.25 dans nos expériences, car elle s'est avérée être suffisante pour capturer l'ensemble des pixels du contour de la partie tête-épaules pour la plupart des postures et angles de vue du corps humain. Notons que, comme la valeur du paramètre n est choisie en fonction de la longueur totale du contour de la silhouette, la longueur du contour tête-épaules candidat extrait est invariante aux changements d'échelle.
Le résultat de l'extraction du contour tête-épaules candidat pour le point saillant "A" sur la Figure 2.10 est mis en évidence par la couleur rouge.

2.3.3.2.2. Extraction des caractéristiques de forme

Après l'extraction des contours candidats, l'étape suivante consiste à extraire leurs caractéristiques de forme. Ces caractéristiques seront utilisées plus tard pour détecter la présence ou non de la partie tête-épaules humaine au sein de la silhouette de l'objet en mouvement. Dans la littérature, il existe plusieurs descripteurs pour décrire la forme d'un contour (Yang et al., 2008), parmi lesquels on peut citer le descripteur de Fourier, le contexte de forme, l'espace de courbure multi-échelle, etc. Dans notre approche, nous avons adopté le descripteur Histogramme de Chaîne de Codes (Chain Code Histogram, CCH) (C. Wang et al., 2012), en raison de sa simplicité et sa faible consommation en termes de temps de calcul. Ce descripteur est basé sur la capture de l'information directionnelle du contour à décrire à l'aide d'une représentation par une Chaîne de Code (CC) (Žalik et al., 2015).

Avant d'illustrer le concept du descripteur CCH, définissons tout d'abord le i-ème contour tête-épaules candidat extrait dans l'étape précédente, en utilisant l'expression 2.26.

$$\Omega_i = \{p_j | j = 1, \dots, l_i\} \tag{2.26}$$

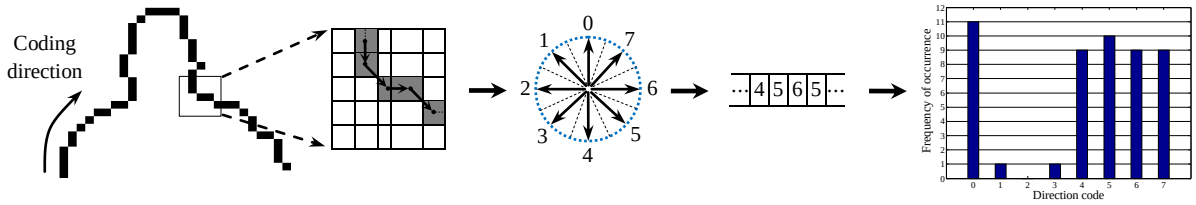
où p_j est le j-ème pixel le long du contour tête-épaules candidat Ω_i, et l_i est sa longueur totale (en pixels). Ensuite, en partant d'un point extrême de Ω_i, la représentation par une CC est obtenue en se déplaçant le long de Ω_i dans le sens des aiguilles d'une montre et en affectant à chaque pixel p_j de Ω_i un code c_j qui indique la direction du pixel suivant p_{j+1} qui appartient à Ω_i. Notons que, dans nos expériences, le codage des directions des pixels le long des contours tête-épaules candidats est effectué en utilisant le descripteur CC à 8 directions montré sur la Figure 2.11(c). Dans ce descripteur, les directions des pixels le long du contour sont codées en utilisant la séquence suivante de huit entiers consécutifs :

$$\Sigma_8 = \{0, 1, 2, 3, 4, 5, 6, 7\} \tag{2.27}$$

Basé sur l'ensemble de codes ci-dessus, pour chaque pixel p_j du contour tête-épaules candidat Ω_i à décrire est assigné un code de direction $c_j \in \Sigma_8$, selon les règles suivantes:

- $c_j = 0$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{3\pi}{8}, \frac{5\pi}{8}]$.
- $c_j = 1$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{5\pi}{8}, \frac{7\pi}{8}]$.
- $c_j = 2$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{7\pi}{8}, \frac{9\pi}{8}]$.
- $c_j = 3$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{9\pi}{8}, \frac{11\pi}{8}]$.
- $c_j = 4$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{11\pi}{8}, \frac{13\pi}{8}]$.
- $c_j = 5$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{13\pi}{8}, \frac{15\pi}{8}]$.
- $c_j = 6$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{15\pi}{8}, \frac{\pi}{8}]$.
- $c_j = 7$ lorsque le prochain vecteur de direction de pixel est dans l'intervalle $[\frac{\pi}{8}, \frac{3\pi}{8}]$.

Un exemple illustrant l'extraction de la CC à 8 directions pour une section d'un contour tête-épaules est montré dans les Figures 2.11(a–d).



(a) (b) (c) (d) (e)

Figure 2.11 : Description d'un contour en utilisant les descripteurs CC et CCH. (a) Exemple d'un contour tête-épaules humain. (b) Section du contour. (c) Descripteur CC à 8 directions. (d) Représentation par une CC. (e) CCH à 8 directions pour l'ensemble du contour tête-épaules.

Parmi les avantages les plus importants de la CC sont sa compacité de représentation, son coût de calcul relativement faible, et son invariance à la translation (ce qui signifie qu'un décalage du contour de la silhouette dans le plan image ne modifie pas la CC). Cependant, ce type de représentation de contour souffre de deux inconvénients majeurs, à savoir la sensibilité aux changements d'échelle (une personne qui se déplace de loin vers la caméra) et aux rotations (un changement de posture ou d'angle de vue de la caméra). En plus de ces inconvénients, le vecteur de caractéristiques généré par la représentation par une CC est souvent de grande dimension, ce qui peut conduire à un processus de classification très coûteux en temps de calcul. Une solution courante pour réduire la dimension de la CC consiste à calculer l'Histogramme de Chaîne de Codes (CCH) (C. Wang et al., 2012).

Le CCH pour un contour est obtenu en calculant les fréquences d'occurrence des différents codes de direction présents dans sa représentation par une CC. Le CCH est typiquement exprimé sous la forme d'un histogramme, dans lequel chaque élément (ou *bin*) est calculé en utilisant l'expression 2.28.

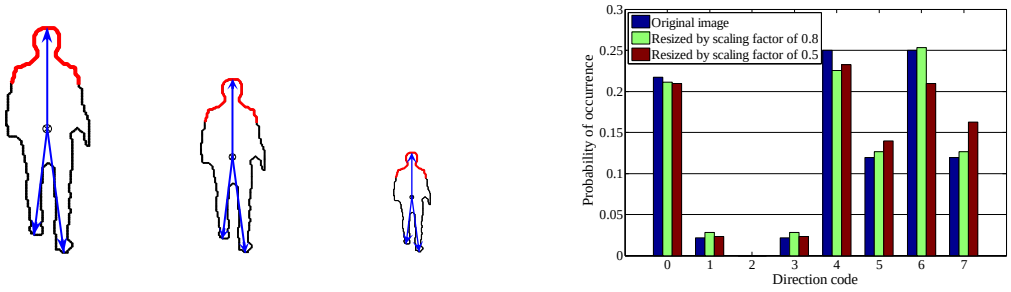
$$\mathrm{CCH}_i(k) = f_k, \qquad k = 0,1,\dots,N_c - 1 \tag{2.28}$$

où f_k est la fréquence d'apparition du code de direction k présent dans la représentation par une CC du contour, et N_c (égal à 8) est le nombre de codes de direction dans le descripteur CC. La représentation par un CCH pour l'exemple de contour tête-épaules candidat donné dans la Figure 2.11(a) est montrée dans la Figure 2.11(e).

Afin de rendre la représentation par un CCH invariante aux changements d'échelle, une solution simple consiste à le normaliser en utilisant l'équation 2.29.

$$\mathrm{NCCH}_i(k) = \frac{\mathrm{CCH}_i(k)}{\sum_{k=0}^{N_c-1} \mathrm{CCH}_i(k)}, \qquad k = 0,1,\ldots,N_c - 1 \tag{2.29}$$

Les *bins* du CCH normalisé (NCCH) fournissent ainsi une estimation de la probabilité d'occurrence des différents codes de direction présents dans la représentation par une CC du contour. A titre d'illustration, un exemple montrant la propriété d'invariance aux changements d'échelle du NCCH est donné à la Figure 2.12.



(a) (b) (c) (d)

Figure 2.12: Propriété d'invariance du CCH normalisé (NCCH) aux changements d'échelle. (a) Exemple d'une silhouette humaine avec le contour tête-épaules candidat mis en évidence en rouge. (b) La silhouette redimensionnée par un facteur d'échelle de 0.8. (c) La silhouette redimensionnée par un facteur d'échelle de 0.5. (d) Comparaison des NCCHs du contour tête-épaules candidat avant et après les changements d'échelle.

Dans cette figure, la Figure 2.12(a) montre un exemple d'une silhouette humaine avec le contour tête-épaules candidat mis en évidence en rouge, les Figures 2.12(b) et 2.12(c) montrent la silhouette redimensionnée par des facteurs d'échelle de 0.8 et 0.5, respectivement, alors que la Figure 2.12(d) montre une comparaison des CCHs normalisés (NCCHs) du contour tête-épaules candidat pour les différentes échelles de la silhouette. À partir de cette dernière figure, nous pouvons observer que la

forme des NCCHs du contour tête-épaules candidat après les changements d'échelle est globalement identique à celle du NCCH de la silhouette originale.

Après avoir résolu le problème de variation aux changements d'échelle du descripteur CCH, le dernier problème restant à résoudre est la variation aux rotations. Ce problème de variation aux rotations du CCH est principalement dû à l'arrangement fixe des codes de direction dans le descripteur CC conventionnel, représenté sur la Figure 2.11(c). Comme ces codes de direction sont disposés de manière circulaire, nous pouvons atteindre l'invariance à la rotation en faisant tourner les codes de direction avec le même angle de rotation et dans le même sens lors de l'extraction de la CC. Cependant, comme le degré et le sens de rotation ne peuvent être connus a priori, nous devons trouver une caractéristique de la silhouette telle que, lorsqu'elle subit une rotation, cette caractéristique devrait également subir une rotation du même degré et dans le même sens. D'autre part, dans nos expériences, nous avons constaté que lorsque la silhouette de l'objet en mouvement détecté subit une rotation dans le sens horaire ou antihoraire (par rapport à son centre de gravité), son squelette-étoile correspondant subit également cette rotation du même degré et dans le même sens de rotation. Ainsi, sur la base de cette observation, et afin de rendre le NCCH invariant aux rotations de la silhouette, nous proposons de faire tourner les codes de direction du descripteur CC conventionnel par rapport à une direction de référence, que nous choisissons comme étant la direction de la branche du squelette-étoile correspondante au contour tête-épaules candidat à décrire.

Afin d'illustrer notre procédure, considérons l'exemple du contour tête-épaules candidat mis en évidence en rouge dans la Figure 2.10. Dans une première étape, le code de direction de la branche (G, A) du squelette-étoile de ce contour est déterminé sur la base du descripteur CC conventionnel montré à la Figure 2.11(c). Dans cet exemple, ce code de direction est trouvé égal à "7", comme le montre la Figure 2.13(a). Dans une deuxième étape, la direction déterminée est prise comme une référence et tous les codes de direction du descripteur CC conventionnel sont tournés, avec un pas de 1, dans le sens inverse des aiguilles d'une montre jusqu'à ce que le code affecté à la direction de référence prenne la valeur "0". Dans notre exemple, cette opération aboutit au nouvel arrangement de codes illustré dans la Figure 2.13(b). Finalement, sur la base de ce nouveau descripteur, que nous dénommons RCC (Rotated Chain Code), le CCH normalisé (NCCH) défini dans

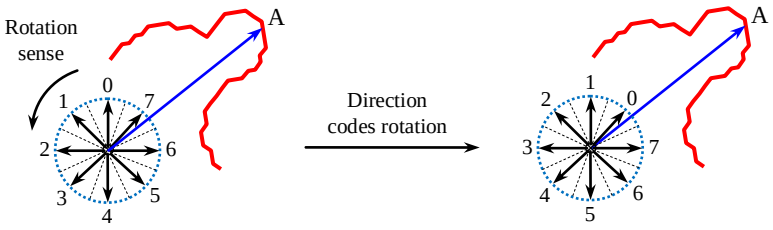
l'équation 2.29, qui est maintenant invariant aux rotations, peut être calculé et utilisé comme une caractéristique pour décrire le contour tête-épaules candidat. Nous pouvons ainsi exprimer la formule générale pour le descripteur RCC comme suit :

$$\Sigma^{\Omega_i} = \left\{ mod\left(\left(k - \left(N_c - D_{\Omega_i} \right) \right), N_c \right) \middle| k = 0, 1, \dots, N_c - 1 \right\} \tag{2.30}$$

où D_{Ω_i} est le code de direction, basé sur le descripteur CC conventionnel, de la branche du squelette-étoile du contour tête-épaules candidat Ω_i à décrire, N_c (égal à 8), le nombre de codes de direction, *mod* est l'opérateur modulo, et Σ^{Ω_i}, le nouvel arrangement des codes de direction dans le descripteur RCC.

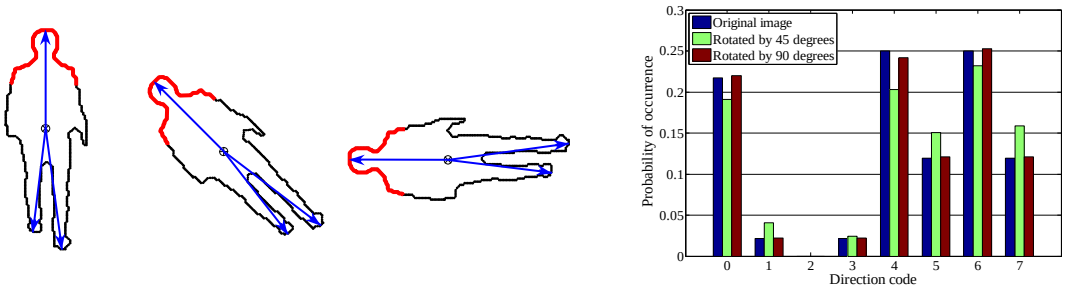
Comme nous pouvons le voir à partir de la formule de l'équation 2.30, l'arrangement des codes de direction dans le descripteur RCC dépend de la direction D_{Ω_i} de la branche du squelette-étoile du contour tête-épaules candidat à décrire. Sur la base de cette direction, l'opérateur *mod* effectue un décalage circulaire à gauche de la séquence de code d'origine (équation 2.27) par un nombre de positions égal à D_{Ω_i}, de telle sorte que l'ensemble des codes de direction reste le même. Ce décalage circulaire entraîne un NCCH invariant aux rotations, car maintenant l'arrangement des codes de direction du descripteur CC dépend de la direction de la branche du squelette-étoile du contour tête-épaules candidat à décrite et non pas d'un arrangement fixe comme cela est montré à la Figure 2.11(c).

A titre d'illustration, un exemple montrant la propriété d'invariance aux rotations du NCCH basé sur le descripteur RCC est donné à la Figure 2.14. Dans cette figure, la Figure 2.14(a) montre une silhouette humaine avec le contour tête-épaules candidat mis en évidence en rouge, les Figures 2.14(b) et 2.14(c) montrent, respectivement, la silhouette tournée de 45 et 90 degrés dans le sens inverse des aiguilles d'une montre, et la Figure 2.14(d) montre une comparaison des NCCHs du contour tête-épaules candidat pour les différentes rotations de la silhouette. À partir de la Figure 2.14(d), nous pouvons observer que la forme globale des NCCHs du contour tête-épaules candidat après avoir effectué les rotations ne change pas trop par rapport à celle avant les rotations. Les petites erreurs dans les NCCHs sont principalement dues à la nature numérique des images prises en compte.



(a) (b)

Figure 2.13: Descripteur RCC (Rotated Chain Code).



(a) (b) (c) (d)

Figure 2.14: Propriété d'invariance aux rotations du CCH normalisé (NCCH). (a) Exemple d'une silhouette humaine avec le contour tête-épaules candidat mis en évidence en rouge. (b) La silhouette soumise à une rotation de 45° dans le sens antihoraire. (c) La silhouette soumise à une rotation de 90° dans le sens antihoraire. (d) Comparaison des NCCHs du contour tête-épaules candidat avant et après les rotations.

2.3.3.2.3. Support Vector Machines (SVM)

Après avoir extrait les caractéristiques des contours tête-épaules candidats, l'étape suivante consiste à les introduire dans un algorithme de Support Vector Machine (SVM) pour détecter éventuellement la présence de la partie tête-épaules humaine au sein de la silhouette de l'objet en mouvement.

L'algorithme SVM (Fletcher, 2009) est une technique d'apprentissage statistique qui est particulièrement adaptée aux problèmes de classification binaire. En raison de ses performances excellentes, cette technique a été appliquée avec succès dans de nombreux problèmes de reconnaissance de formes, tels que la détection d'objets, la reconnaissance de caractères manuscrits, la vérification de la parole et du locuteur,

la recherche d'images par leur contenu, etc. L'objectif principal de l'algorithme SVM est de trouver un hyperplan optimal qui partitionne les échantillons de l'ensemble de données d'apprentissage en deux classes, "positive" et "négative", de telle sorte que la marge entre l'hyperplan et les échantillons d'apprentissage les plus proches (appelés "vecteurs de support") dans chaque classe est maximisée.

Ainsi, soit un ensemble de données d'apprentissage $\{(x_i, y_i),\ i = 1, \dots, \ell\}$ de ℓ échantillons, avec $x_i \in \mathbb{R}^n$ est le vecteur de caractéristiques de dimension n du i-ème échantillon, et $y_i \in \{+1, -1\}$ est la classe à laquelle il appartient.

Pour des données linéairement séparables, l'hyperplan séparateur optimal peut être défini par l'expression suivante :

$$f(x) = w \cdot x + b = \sum_{i=1}^{\ell} w_i x_i + b = 0 \tag{2.31}$$

où, $w \in \mathbb{R}^d$ est un vecteur de poids de la même dimension que l'espace des caractéristiques, et b est un biais. Ces deux paramètres déterminent la position de l'hyperplan séparateur dans l'espace des caractéristiques, et ils sont sélectionnés de manière à obéir aux contraintes suivantes (Cortes and Vapnik, 1995):

$$y_i(w \cdot x_i + b) - 1 \geq 0 \Leftrightarrow \begin{cases} f(x_i) = w \cdot x_i + b \geq 1, & si\ \ y_i = +1 \\ f(x_i) = w \cdot x_i + b \leq -1, & si\ \ y_i = -1 \end{cases} \tag{2.32}$$

Cependant, lorsque les échantillons de données sont linéairement non séparables, la solution pour le classifieur SVM peut être obtenue en utilisant une projection non linéaire des échantillons de données vers un espace de caractéristiques de plus grande dimension dans lequel les échantillons peuvent être linéairement séparables. L'hyperplan séparateur optimal peut ainsi être déterminé en résolvant le problème d'optimisation suivant :

$$\begin{aligned} min \quad & \Phi(w) = \frac{1}{2}(w \cdot w) + \mathrm{C} \sum_{i=1}^{\ell} \xi_i \\ s.t. \quad & y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i, i = 1, \dots, \ell \end{aligned} \tag{2.33}$$

où, $\xi_i \geq 0$ sont des variables "ressorts" (slack variables) qui permettent à un exemple d'être dans la marge $0 \leq \xi_i \leq 1$ (appelée aussi "erreur de marge"), C est le paramètre de régularisation qui contrôle le compromis entre la complexité de la fonction de décision et les erreurs de classification, et $\phi(x_i)$ est la fonction qui

permet de projeter l'échantillon de données x_i de l'espace de caractéristiques d'entrée vers un espace de plus grande dimension. Le problème d'optimisation de l'équation 2.33 peut être simplifié en le transformant avec la condition de Kunh-Tucker en un problème dual lagrangien équivalent comme suit :

$$\begin{aligned} max \;\; & W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ s.t. \; & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \qquad 0 \leq \alpha_i \leq C, i = 1, \dots, \ell \end{aligned} \tag{2.34}$$

où $K(x_i, x_j)$ est le noyau qui, selon le théorème de Mercer (Cortes and Vapnik, 1995), est défini comme une fonction telle que $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. Les noyaux les plus couramment utilisés dans la pratique sont les suivants:

- *Linéaire*: $K(x_i, x) = (x \cdot x_i)$
- *Polynomial*: $K(x_i, x) = \big((x \cdot x_i) + 1\big)^d$, avec d est l'ordre du noyau.
- *Fonction* de *Base Radiale (RBF)*: $K(x_i, x) = e^{\left(-\frac{\|x - x_i\|^2}{2\sigma_{rbf}^2}\right)}$, avec σ_{rbf} est la largeur du noyau.
- *Perceptron Multicouche*: $K(x_i, x) = \tanh(\gamma(x \cdot x_i) + \delta)$, avec γ (un nombre positif) et δ (un nombre négatif) sont les paramètres du noyau.

Dans nos expériences, nous avons testé tous ces différents noyaux avec différentes valeurs pour leurs paramètres, et celui qui a atteint les meilleures performances a été adopté.

Après avoir obtenu la solution α^* pour le problème de l'équation 2.34, les deux paramètres w^*et b^*de l'hyperplan séparateur optimal peuvent être alors calculés par les expressions suivantes :

$$w^* = \sum_{i=1}^{\ell} y_i \alpha_i^* \phi(x_i) \tag{2.35}$$

$$\alpha_i^* (y_i (w^* \cdot \phi(x_i) + b^*) - 1) = 0 \tag{2.36}$$

La classification d'un échantillon de données test $\boldsymbol{x}$ peut être effectuée en utilisant la fonction de décision suivante:

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{\ell} \alpha_i^* y_i K(x_i, \boldsymbol{x}) + b^*\right) \tag{2.37}$$

Après l'application de l'algorithme SVM sur tous les contours tête-épaules candidats extraits, la décision est prise comme suit. Si aucun de ces contours n'est reconnu comme étant une partie tête-épaules humaine, notre méthode décide que la silhouette détectée provient d'un objet non-humain en mouvement. Inversement, si l'un des contours candidats est reconnu comme étant une partie tête-épaules humaine, alors notre méthode décide que l'objet détecté peut probablement être un humain. Dans ce cas, un deuxième test est effectué afin de détecter la présence ou non des deux jambes humaines au sein de la silhouette de l'objet en mouvement. La procédure pour effectuer cette tâche est décrite dans la sous-section suivante.

2.3.3.3. Détection des jambes

Les jambes, généralement visibles sous de nombreux angles de vue, sont parmi les parties qui distinguent un corps humain d'autres objets non humains en mouvement. Cependant, dans des situations réelles de vidéo surveillance, la détection des jambes humaines est une tâche très difficile à atteindre en raison de la grande variation de l'apparence du corps humain durant son mouvement. L'exemple de la Figure 2.9, vue d'un angle latéral, la silhouette humaine avec des jambes fermées apparaît très différente d'une silhouette avec des jambes complètement séparées. De plus, la position relative des jambes humaines par rapport à l'emplacement de la tête peut varier considérablement en cas de présence de changements sévères de la posture et/ou de l'angle de vue du corps humain. Ainsi, dans notre approche, afin de traiter ces problèmes particuliers et réduire le nombre de fausses positives, nous proposons une procédure de détection des jambes humaines très similaire à celle adoptée dans la première approche (sous-section 2.3.2.1). A la différence de la première procédure, qui suppose que les personnes en mouvement se déplacent dans la posture "debout", cette deuxième procédure est adaptative par rapport aux changements de posture et d'angle de vue, et elle repose sur l'utilisation de la position de la partie tête-épaules humaine (détectée dans l'étape précédente) et les deux caractéristiques décrites dans les sous-sections 2.3.1.4 et 2.3.3.1, à savoir le squelette-étoile et la meilleure ellipse

ajustée. Les trois étapes principales de cette procédure de détection des jambes humaines peuvent être résumées ainsi :

Dans la première étape, il s'agit d'identifier la partie (supérieure ou inférieure) de la silhouette contenant la forme tête-épaules humaine détectée, puis choisir sa partie opposée comme une zone pour détecter les jambes.

Dans la deuxième étape, on définit un intervalle de recherche $[-\theta_{legs}, +\theta_{legs}]$ à l'intérieur de la partie opposée, en formant un angle de θ_{legs} (dont la valeur "optimale" est trouvée empiriquement) degrés dans les deux directions horaire et antihoraire à partir de l'axe semi-majeur de la meilleure ellipse ajustée situé dans la partie opposée.

Dans la troisième étape, on compte le nombre de branches du squelette-étoile situées à l'intérieur de l'intervalle $[-\theta_{legs}, +\theta_{legs}]$. Si ce nombre est égal à 1 (ce qui correspond à des jambes humaines complètement fermées) ou 2 (ce qui correspond à des jambes humaines complètement séparées), alors, considérer l'objet en mouvement détecté comme étant un humain. Inversement, si aucune branche du squelette-étoile n'est trouvée à l'intérieur de la région de recherche, ou si le nombre trouvé est supérieur à 2, alors, considérer l'objet en mouvement détecté comme étant une fausse alarme et rejeter-le.

A titre d'illustration, quelques exemples montrant la procédure décrite ci-dessus lorsqu'elle est appliquée sur des silhouettes humaines en différentes postures sont donnés dans la Figure 2.15. Sur cette figure, les contours mis en évidence en rouge représentent les parties tête-épaules humaines détectées, alors que les secteurs noirs ombrés représentent l'intervalle prédéfini $[-\theta_{legs}, +\theta_{legs}]$ pour détecter les jambes. Comme nous pouvons l'observer à partir de cette figure, notre procédure proposée peut détecter avec succès les jambes humaines, même en cas de présence de situations difficiles, telles que des changements dans la posture du corps humain.

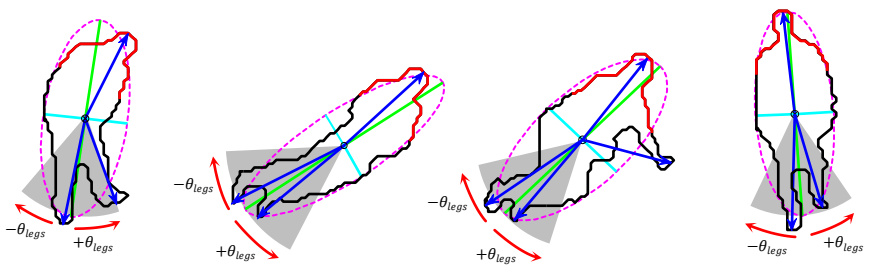


Figure 2.15: Illustration de la procédure pour détecter les jambes humaines.

2.4. Conclusion

Dans ce chapitre, nous avons tout d'abord présenté un état de l'art des méthodes proposées dans la littérature pour la détection de personnes dans des images et des séquences d'images IR. Deux principales approches ont été détaillées, à savoir les approches basées sur des caméras proche IR, et les approches basées sur des caméras IR thermiques. Nous avons, ensuite, présenté les deux approches que nous proposons. La première approche consiste à calculer plusieurs sous-fonctions de similarité basées sur des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement, puis de les combiner dans une fonction globale qui a été utilisée pour la détection des personnes. Cette première approche peut être appliquée sur des séquences d'images acquises par des caméras IR thermiques, et malgré sa simplicité, elle a comme avantage de détecter des personnes en mouvement sans la nécessité d'un apprentissage a priori d'un classifieur. La deuxième approche que nous avons proposé consiste à la détection des deux parties qui caractérisent le corps humain, à savoir l'ensemble tête-épaules (ressemblant à la forme de la lettre grecque Omega Ω), et les deux jambes. La détection de la partie tête-épaules a été effectuée en trois étapes principales. Dans la première étape, les contours tête-épaules candidats de la silhouette de l'objet en mouvement sont extraits en utilisant sa meilleure ellipse ajustée et son squelette-étoile. Ensuite, dans la deuxième étape, ces contours candidats sont passés à travers une phase d'extraction des caractéristiques afin d'extraire leurs caractéristiques de forme, qui sont dans notre cas les histogrammes de chaîne de codes normalisés (NCCHs). Enfin, dans la dernière étape, ces caractéristiques sont introduites dans un classifieur SVM entraîné pour détecter la présence de la forme tête-épaules humaine au sein de la silhouette de l'objet en mouvement. Quant à la détection des jambes, elle est effectuée par la vérification de la présence d'un ou de deux segments de ligne dans la partie du squelette-étoile opposée à celle contenant la forme tête-épaules détectée. Comme cette deuxième approche proposée est basée uniquement sur la silhouette des objets en mouvement, elle peut être appliquée sur des séquences d'images IR acquises par des caméras thermiques ou par des caméras proche IR. Et contrairement à la plupart des méthodes d'état de l'art, cette approche a comme avantage de détecter des personnes même en cas de présence de changements dans leur posture. Dans

le prochain chapitre, nous présentons la deuxième partie de notre système proposé, à savoir le suivi de personnes en mouvement dans des séquences d'images IR.

Chapitre 3

Suivi de personnes

3.1. Introduction

Le suivi des personnes dans des séquences d'images IR constitue une tâche très importante dans de nombreux domaines parmi lesquels on peut citer la vidéo surveillance intelligente. Dans tous les cas d'application, l'algorithme de suivi devra être suffisamment robuste pour pouvoir gérer des situations difficiles telles que les clutters, les changements d'apparence, la non-rigidité du corps humain et la présence d'objets non-humains dans la scène surveillée. En plus de ces difficultés majeures, l'arrière-plan des images IR est souvent affecté par du bruit, et le contraste entre les objets ou personnes en mouvement et l'arrière-plan est très faible. De plus, dans les cas réels de vidéo surveillance, le suivi est effectué en même temps sur plusieurs personnes dans la scène surveillée, ce qui augmente la probabilité d'apparition des occultations partielles ou totales entre ces personnes. A cause de toutes ces contraintes, le développement d'un algorithme robuste de suivi de personnes dans des séquences d'images IR reste un grand défi à relever.

Ainsi, de nombreux chercheurs se sont intéressés au suivi d'objets cibles dans des séquences d'images IR, ce qui a produit plusieurs approches et algorithmes dans ce domaine. Parmi les plus populaires, on peut citer la méthode Mean-Shift (Oshima et al., 2006; Wang et al., 2009; Liu and Yang, 2012; Shuang and Yu-Ping, 2012; Yun and Kim, 2019), le filtre de Kalman (Binelli et al., 2005; Xu et al., 2005; Lee et al., 2012; Bhusal, 2015; Shahzad and Jalal, 2022), le filtre à particules (Li and Gong, 2010; Wang and Tang, 2010; Wang et al., 2012; Portmann et al., 2014; Younsi et al., 2020), le flot optique (Bhusal, 2015), le filtre de corrélation (He et al., 2015;

Asha and Narasimhadhan, 2017; Chen et al., 2020), l'appariement de modèles (Bal and Alam, 2004; Sahani et al., 2011; Paravati and Esposito, 2014), et les algorithmes à base de réseaux de neurones convolutifs (CNN) (Q. Liu et al., 2017; Kwan et al., 2019; Xu et al., 2021). Parmi tous ces algorithmes, le filtre à particules est sans doute l'une des approches les plus largement utilisées en raison de ses performances exceptionnelles dans le cas de modèles hautement non linéaires avec un bruit non gaussien. Dans ce sens, la tâche de suivi de cible peut être considérée comme un problème d'estimation bayésienne récursive dont les états estimés représentent certains paramètres de cette cible, tels que sa position ou sa vitesse. Le filtre à particules est basé sur deux modèles de base, l'un est dit de transition des états et l'autre d'observation. Le premier modèle est utilisé pour prédire les positions possibles à l'instant suivant de la cible et le second pour déterminer l'emplacement le plus probable parmi ces positions prédites. Ce dernier modèle influence considérablement la performance du suivi. Il est nécessaire, dans ce cas, pour un algorithme efficace de suivi, de construire un modèle d'observation robuste pour la cible suivie. Dans l'algorithme traditionnel de filtrage à particules (Pérez et al., 2002; Nummiaro et al., 2003) pour le suivi de cibles, le modèle d'observation est construit en se basant uniquement sur une seule caractéristique, à savoir l'intensité, étant donné que celle-ci n'est pas limitée à un type particulier de cible à suivre. Bien qu'il soit parfois satisfaisant pour certaines séquences d'images, cet algorithme traditionnel ne permet pas de gérer certaines situations difficiles rencontrées dans le cas des séquences d'images IR, notamment lorsque les cibles suivies subissent des changements soudains d'apparence ou lorsqu'il s'agit d'un suivi simultané de plusieurs cibles proches ayant une intensité similaire. Heureusement, le filtre à particules a une propriété qui lui permet l'intégration de différentes caractéristiques dans le modèle d'observation. Ainsi, afin d'accroître la robustesse et la fiabilité de l'algorithme de suivi, il est intéressant de combiner plusieurs caractéristiques.

Dans ce chapitre, nous présentons une nouvelle approche pour le suivi de personnes dans des séquences d'images IR en utilisant un filtre à particules et une combinaison adaptative d'informations provenant de plusieurs types de caractéristiques dont l'intensité, la texture, la vélocité de mouvement et la distance spatiale. Dans un premier temps, plusieurs modèles dans ces différentes caractéristiques sont créés en ligne, puis sont ultérieurement mis à jour afin de les

adapter aux changements significatifs de l'apparence de l'être humain détecté. Lorsque les nouvelles observations arrivent avec les trames suivantes, les distances de similarité entre les différents modèles créés et les régions en mouvement observées sont calculées. Ces distances de similarité individuelles sont enfin combinées, dans le cadre d'un filtre à particules, en utilisant des poids adaptatifs afin de suivre l'être humain détecté de manière efficace. Pour augmenter davantage la robustesse de notre méthode de suivi, nous introduisons aussi une stratégie automatique de détection et de traitement des occlusions basée sur des règles heuristiques simples ainsi que l'histogramme de projection verticale (VPH) en niveaux de gris.

Le reste de ce chapitre est organisé comme suit. Tout d'abord, dans la section 3.2, nous présentons une revue des méthodes proposées dans la littérature pour le suivi de personnes dans des séquences d'images IR. Ensuite, dans la section 3.3, nous présentons un aperçu général de l'algorithme de filtrage à particules. Dans la section 3.4, nous présentons la nouvelle approche que nous avons développée pour le suivi de personnes, et qui est basée sur le filtrage à particules et une combinaison adaptative d'informations provenant de plusieurs types de caractéristiques. Enfin, dans la section 3.5, nous terminons ce chapitre par une conclusion qui résume l'essentiel de nos contributions.

3.2. Revue des méthodes de suivi de personnes dans des séquences d'images IR

Considéré comme l'un des principaux sujets de recherche en vision par ordinateur, le suivi d'objets peut être défini comme la tâche d'estimation, ou de génération, du chemin et de la trajectoire d'un objet dans le plan de l'image, et ce, en localisant sa position dans chaque trame de la séquence d'images (Joshi et al., 2018). Le suivi des personnes dans des séquences d'images IR est considéré comme étant un cas particulier du problème général de suivi visuel d'objets. Afin d'atteindre une meilleure robustesse et de réduire l'incertitude dans le suivi de personnes dans des séquences d'images IR, de nombreux efforts de recherche ont été menés ces dernières années dans ce domaine pour améliorer les performances du suivi en concevant divers algorithmes et en utilisant différentes caractéristiques. Les approches basées sur des caractéristiques d'apparence, notamment l'intensité, sont les méthodes les plus conventionnelles et les plus largement utilisées dans la

littérature. Yasuno et al., (2004), par exemple, ont extrait la région de la tête (région de haute intensité) de l'humain détecté, puis ils l'ont utilisée comme un modèle de correspondance pour suivre l'humain d'une trame à une autre en utilisant une simple procédure de prédiction de premier ordre. Xu et al., (2005) ont utilisé le filtre de Kalman pour prédire la position approximative de l'humain détecté, puis ils ont appliqué l'algorithme de mean-shift autour de cette position prédite pour localiser l'humain avec précision. Kumar, (2013) a utilisé un algorithme de suivi qui crée une carte de confiance dans la nouvelle trame basée sur l'histogramme de densité de l'humain détecté dans la trame précédente, puis il cherche le pic de cette carte de confiance autour la position précédente de l'humain. L'avantage des approches basées sur l'intensité est qu'elles sont applicables à de multiples conditions environnementales et peuvent détecter des personnes à des échelles variables. Cependant, leur inconvénient est que leur performance peut se dégrader si l'intensité de l'humain suivi est similaire à celle de l'arrière-plan. Afin de rendre le suivi plus robuste et plus fiable, d'autres approches basées sur des caractéristiques de texture comme l'entropie des ondelettes (Li and Wang, 2009), ou de forme, telles que le descripteur SURF (Jüngling and Arens, 2010), et le descripteur HOG (Kim and Kwon, 2015) ont été proposées. Comme ces approches sont basées sur les variations d'intensité plutôt que sur les intensités, elles sont très stables et robustes dans diverses conditions environnementales. Cependant, l'utilisation de techniques très coûteuses en termes de temps de calcul rend ces approches moins appropriées pour des traitements en temps réel. Afin d'atteindre encore plus de robustesse, notamment face au bruit et aux encombrements de l'arrière-plan, d'autres approches basées sur des caractéristiques de mouvement ont été proposées. Ran et al., (2007), par exemple, ont présenté une approche pour la détection et le suivi de personnes basée sur l'estimation de la périodicité de la marche humaine. Les auteurs estiment la fréquence des mouvements périodiques du corps humain en utilisant un test d'hypothèse à deux étapes en cascade pour filtrer les pixels non périodiques (ou stationnaires) de telle sorte à fonctionner de manière optimale pour les directions de marche radiale et latérale. Cette approche est très robuste contre les mouvements de la caméra, le bruit du capteur et les arrière-plans encombrés, mais son inconvénient est qu'elle nécessite un bon alignement des trames, car elle utilise l'information temporelle à l'échelle du pixel. Portmann et al., (2014) ont proposé un algorithme de suivi basé sur un filtre à

particules qui combine à la fois les résultats de la détection humaine et des contraintes temporelles pour augmenter la précision de l'identification et de la localisation des personnes. Plus récemment, Yang et al., (2017) ont proposé un algorithme spatio-temporel pour le suivi humain dans des séquences d'images IR en utilisant des volumes de tranches horizontales et verticales (horizontal and vertical slice volumes) pour obtenir des tubulures de trajectoire (trajectory manifolds). En analysant les variations de la largeur et de la longueur de ces tubulures de trajectoire, la boîte englobante et la position de la cible sont obtenues puis utilisées pour réaliser la tâche de suivi. Cette approche peut gérer les occultations partielles ou complètes, mais elle n'est pas robuste pour le suivi de cibles multiples. De plus, elle n'est applicable que pour des scénarios hors-ligne.

Notons que la plupart des approches décrites précédemment sont basées sur un seul type de caractéristiques déterminées a priori de manière aléatoire, ou parfois, en effectuant des tests préliminaires sur des bases de données représentatives. Cependant, l'utilisation d'un seul type de caractéristiques pour le suivi humain s'est avérée insuffisante pour faire face avec succès à la grande varieté de situations rencontrées dans des scenarios du monde réel. Pour surmonter ce problème et améliorer davantage la robustesse du suivi de personnes, les méthodes basées sur de multiples caractéristiques ont attiré l'attention des chercheurs et un certain nombre d'articles scientifiques ont été publiés dans la littérature (Oshima et al., 2006; Li and Gong, 2010; He et al., 2015; Gao and Jhang, 2016; Wang et al., 2019). Oshima et al., (2006), par exemple, ont proposé une méthode pour le suivi d'une seule personne en utilisant une caméra statique proche IR. Le suivi est effectué en utilisant la méthode de mean-shift combinée avec trois histogrammes différents, basés sur l'amplitude du flot optique, la direction du flot optique et la couleur. Li and Gong, (2010) ont construit un histogramme des régions d'intérêt dans un espace de projection intensité-distance pour surmonter l'inconvénient de l'insuffisance des informations lorsque seule la caractéristique d'intensité est prise en compte. Cet histogramme est ensuite introduit dans un filtre à particules pour réaliser un suivi robuste. He et al., (2015) ont présenté un algorithme de suivi basé sur la représentation de rang faible (low-rank representation) et un filtre de corrélation pondéré, qui intègre une fonction de pondération basée sur trois caractéristiques différentes, à savoir une caractéristique d'intensité, une caractéristique spatiale et une caractéristique de mouvement. Gao and Jhang,

(2016) ont proposé une méthode de suivi de personnes dans des séquences d'images IR basée sur la représentation éparse (sparse representation) et une combinaison de trois caractéristiques multiples qui sont l'histogramme d'intensité, l'entropie locale et la différence de moyennes de contraste locale. Ding et al., (2022) ont combiné des caractéristiques convolutives extraites en utilisant un CNN pré-entraîné et des caractéristiques HOG pour représenter la cible suivie dans le cadre du filtre de corrélation.

Toutes les approches citées dans le paragraphe précédent combinent les caractéristiques de manière non adaptative, ce qui signifie que la fiabilité de chaque caractéristique est supposée être inchangée durant le suivi. Cependant, une telle hypothèse est souvent invalide dans les cas réels en raison des changements dynamiques de la scène surveillée. Pour remédier à ce problème, d'autres approches de suivi basées sur une combinaison adaptative de caractéristiques ont été proposées (Dai et al., 2007; Wang et al., 2009; Wang and Tang, 2010; J. Wang et al., 2012; Asha and Narasimhadhan, 2017; Yu et al., 2019). Dai et al., (2007) ont employé une méthode basée sur l'appariement de graphes (graph matching), qui exploite conjointement les informations de forme, d'apparence et de distance. La combinaison de ces caractéristiques est effectuée adaptativement sur la base du taux de chevauchement entre les boîtes englobantes des humains détectés. Wang et al., (2009) ont extrait les caractéristiques d'intensité et de contours pour améliorer la performance de l'algorithme Mean-shift, puis ils ont utilisé les informations de mouvement pour guider ces caractéristiques fusionnées durant le processus de suivi humain. Les caractéristiques d'intensité et de contours sont également combinées de manière adaptative par Wang and Tang, (2010), et Wang et al., (2012) pour effectuer le suivi humain dans un cadre de filtre à particules. Asha and Narasimhadhan, (2017) ont proposé une méthode de suivi de personnes basée sur un filtre de corrélation à noyau (Kernelized Correlation Filter, KCF) et une combinaison adaptative de caractéristiques d'amplitude de gradients et d'histogrammes d'intensité spatiale. Yu et al., (2019) ont également utilisé les filtres de corrélation avec une fusion adaptative de plusieurs caractéristiques, incluant l'amplitude de gradients, une caractéristique de mouvement, FHOG (Felzenszwalb's Histogram of Oriented Gradient), et une caractéristique d'intensité. Plus récemment, et afin de mieux modéliser l'apparence des cibles à suivre, Yuan et al., (2023) ont proposé de fusionner, de manière adaptative, plusieurs types de

caractéristiques, y compris des caractéristiques extraites manuellement (handcrafed features) et des caractéristiques profondes extraites par des réseaux de neurones convolutifs.

En nous inspirant des méthodes décrites ci-dessus, dans ce chapitre, nous proposons une nouvelle méthode pour le suivi de personnes dans des séquences d'images IR en utilisant un filtre à particules et une combinaison d'informations provenant de plusieurs types de caractéristiques, à savoir l'intensité, la texture, la vélocité de mouvement et la distance spatiale. La combinaison de ces caractéristiques est effectuée de manière adaptative en tenant compte de la capacité discriminative de chaque caractéristique durant le processus de suivi.

3.3. Aperçu général de l'algorithme de filtrage à particules

Le filtre à particules (Isard and Blake, 1998; Pérez et al., 2002; Nummiaro et al., 2003) est un algorithme de filtrage qui a été développé sur la base de la méthode de Monte-Carlo séquentielle et l'estimation bayésienne récursive d'état. L'idée de base de cet algorithme est d'utiliser un ensemble d'échantillons aléatoires, avec des poids associés, pour estimer récursivement la fonction de densité de probabilité (PDF) a *posteriori* de l'état d'un système à partir de mesures obtenues en ligne. En raison de ses performances supérieures par rapport à des méthodes conventionnelles, telles que le filtre de Kalman (Kim and Bang, 2018), l'algorithme de filtrage à particules a été largement utilisé durant ces dernières années pour résoudre le problème de suivi d'objets dans le cas de modèles hautement non linéaires et non gaussiens.

Dans le contexte du suivi visuel, supposons que l'état, à l'instant k, de l'objet à suivre est noté par $\mathbf{x}_k$, et que la séquence de toutes les observations disponibles jusqu'à l'instant k est notée par $\mathbf{z}_{1:k} = \{\mathrm{z}_1, \mathrm{z}_2, \dots, \mathrm{z}_\mathrm{k}\}$. Le filtre à particules consiste à résoudre le problème de suivi en se basant sur le modèle de mouvement, défini dans l'équation 3.1, et sur le modèle d'observation, défini dans l'équation 3.2. Le premier modèle décrit l'évolution de l'état en un pas de temps, tandis que le second décrit la relation entre l'état du système et les observations au même instant.

$$\mathbf{x}_k = f_k(\mathbf{x}_{k-1},\ \mathbf{u}_k) \tag{3.1}$$

$$\mathbf{z}_k = h_k(\mathbf{x}_k,\ \mathbf{v}_k) \tag{3.2}$$

Dans les équations ci-dessus, f_k et h_k sont des fonctions, éventuellement non linéaires et dépendantes du temps, $\mathbf{u}_k$ et $\mathbf{v}_k$ sont, respectivement, les bruits du système et d'observation, qui sont tous les deux supposés être indépendants avec une fonction de distribution connue. Le problème d'estimation de l'état $\mathbf{x}_k$ dans le cadre d'estimation bayésienne est résolu en calculant la PDF a *posteriori* $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. Ce calcul est effectué en deux étapes récursives, à savoir une étape de prédiction (ou de propagation), exprimée par l'équation 3.3, et une étape de correction (ou de mise à jour), exprimée par l'équation 3.4 :

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})\, d\mathbf{x}_{k-1} \tag{3.3}$$

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{\int p(\mathbf{z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{z}_{1:k-1})\, d\mathbf{x}_k} \tag{3.4}$$

où, $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ est la probabilité de transition d'état de l'instant $k-1$ à l'instant k, et $p(\mathbf{z}_k|\mathbf{x}_k)$ est la PDF des observations arrivées à l'instant k.

Cependant, comme il n'existe souvent pas de solution analytique pour les équations 3.3 et 3.4 en raison de la présence de non-linéarités et de bruit non gaussien dans la plupart des problèmes réels d'estimation d'état, l'algorithme de filtrage à particules approxime la densité a *posteriori* $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ par une somme de N_s fonctions de Dirac (ou particules) centrées en $\{\mathbf{x}_k^{(i)}\}_{i=1}^{N_s}$, comme le montre l'équation suivante :

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} \boldsymbol{w}_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}) \tag{3.5}$$

où $\{\boldsymbol{w}_k^{(i)},\ i = 1, \dots, N_s\}$ sont les poids associés aux particules $\{\mathbf{x}_k^{(i)},\ i = 1, \dots, N_s\}$, et qui sont mis à jour récursivement comme suit (Zhang et al., 2017):

$$\boldsymbol{w}_k^{(i)} \propto \boldsymbol{w}_{k-1}^{(i)} \frac{p\left(\mathbf{z}_k \middle| \mathbf{x}_k^{(i)}\right) p\left(\mathbf{x}_k^{(i)} \middle| \mathbf{x}_{k-1}^{(i)}\right)}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_{1:k}^{(i)})} \tag{3.6}$$

où, $q(\cdot)$ est la fonction de densité d'importance, qui est généralement choisie comme étant égale à $p\left(\mathbf{x}_k^{(i)} \middle| \mathbf{x}_{k-1}^{(i)}\right)$, ce qui conduit à :

$$\boldsymbol{w}_k^{(i)} \propto \boldsymbol{w}_{k-1}^{(i)}\, p\left(\mathbf{z}_k \middle| \mathbf{x}_k^{(i)}\right) \tag{3.7}$$

Cependant, dans la pratique, afin d'éviter le problème de dégénérescence, c'est-à-dire la convergence de tous les poids des particules, excepté un, vers zéro, une opération de ré-échantillonnage est souvent ajoutée à la fin de l'algorithme de filtrage à particules. Cela se fait en mettant $\boldsymbol{w}_{k-1}^{(i)} = \frac{1}{N_s}$, $\forall i$. Par conséquent, l'équation 3.7 peut être simplifiée comme suit :

$$\boldsymbol{w}_k^{(i)} \propto p\left(\mathbf{z}_k \middle| \mathbf{x}_k^{(i)}\right) \tag{3.8}$$

Ce qui signifie que les poids des particules sont proportionnels à la PDF des nouvelles observations conditionnées par chaque état.

3.4. Approche proposée

La méthode que nous proposons pour le suivi de personnes en mouvement dans des séquences d'images IR se compose principalement des sept étapes : 1) Initialisation du suivi, 2) propagation des particules, 3) pondération des particules, 4) estimation d'état, 5) ré-échantillonnage des particules, 6) mise à jour du modèle, et 7) détection et gestion des occultations. Les détails sur chacune de ces étapes sont présentés dans les sous-sections suivantes.

3.4.1. Initialisation du suivi

L'initialisation du suivi est une tâche très importante dans un système de vidéo surveillance intelligente, car c'est elle qui détermine la cible à suivre par l'algorithme dans la scène surveillée. Ainsi, toute erreur dans cette initialisation peut conduire à un résultat peu satisfaisant, et elle peut même provoquer un échec précoce du processus de suivi. La plupart des travaux existant dans la littérature (Li and Gong, 2010; Wang and Tang, 2010; J. Wang et al., 2012) utilisent une initialisation manuelle en définissant, dans la première trame, un rectangle (ou une boîte englobante) autour de la personne (ou cible) à suivre. Cependant, dans le cas pratique, un système de vidéosurveillance intelligent et efficace ne doit pas être dépendent d'une intervention manuelle d'un opérateur pour fonctionner. Ainsi, dans notre système proposé, nous initialisons l'algorithme de suivi automatiquement, c'est-à-dire sans aucune information préalable sur la taille ni

l'emplacement des personnes à suivre dans la scène surveillée. Pour réaliser cette tâche, nous utilisons l'une des méthodes de détection de personnes que nous avons proposées dans le chapitre précédent.

Ainsi, lorsqu'un objet en mouvement est détecté comme étant un humain, l'initialisation de l'algorithme de suivi peut s'effectuer comme suit. Tout d'abord, la dernière position de cet humain dans la scène est enregistrée, et elle est prise comme étant la position d'initialisation pour l'algorithme de suivi. Ensuite, les modèles d'apparence (histogrammes d'intensité et de texture) et de mouvement (vélocité moyenne), qui seront détaillés dans la Section 3.3.3, et qui permettront le suivi de l'humain détecté tout au long de la séquence d'images sont initialisés. Enfin, un filtre à particules est créé pour cet humain en générant un ensemble de N_s particules $\{\mathbf{x}_k^{(i)} = (x_k^{(i)}, y_k^{(i)}),\ i = 1, \dots, N_s\}$ normalement distribuées autour de sa position d'initialisation, avec $(x_k^{(i)}, y_k^{(i)})$ sont les coordonnées à l'instant k de la i-ième particule du filtre.

3.4.2. Propagation des particules (étape de prédiction)

Après avoir initialisé l'algorithme de suivi, l'étape suivante consiste à propager les particules à partir de l'instant k-1 vers l'instant k. Dans ce travail, cette tâche est réalisée en utilisant le modèle de mouvement à vélocité adaptative donné par l'équation 3.9 ci-dessous :

$$\mathbf{x}_k^{(i)} = A\,\mathbf{x}_{k-1}^{(i)} + \overline{\boldsymbol{v}}_{k-1} + \mathbf{u}_k \tag{3.9}$$

où,

- $\mathbf{x}_k^{(i)} = (x_k^{(i)}, y_k^{(i)})^T$ est le vecteur d'état à l'instant k, avec $x_k^{(i)}$ et $y_k^{(i)}$ sont les coordonnées de la i-ième particule dans le plan image.
- A est la partie déterministe du modèle de mouvement. Elle est égale à une matrice identité de taille 2×2.
- $\overline{\boldsymbol{v}}_{k-1}$ (qui sera défini plus tard dans la sous-section 3.3.3) est la vélocité moyenne de l'humain suivi sur les dernières trames récentes de la séquence d'images.
- $\mathbf{u}_k \sim \mathcal{N}(0, Q)$, est un bruit normalement distribué de moyenne nulle et de matrice de covariance Q convenablement choisie. Il décrit l'incertitude dans le vecteur d'état.

3.4.3. Pondération des particules (étape de mise-à-jour)

Une fois que les particules du filtre sont propagées, leurs poids peuvent être calculés. Pour faire face aux situations du monde réel telles que la présence de bruit et de multiples humains en mouvement dans les séquences d'images IR, nous proposons dans ce travail de combiner, de manière adaptative, des informations provenant de quatre caractéristiques différentes, à savoir la proximité spatiale, l'intensité, la texture et l'information de mouvement. Les détails sur ces différentes caractéristiques sont décrits dans les sous-sections suivantes.

3.4.3.1. Proximité spatiale

L'information de proximité spatiale permet aux particules du filtre de se déplacer vers les régions de mouvement dans la scène observée. Dans ce travail, cette caractéristique est obtenue en calculant la distance Euclidienne entre chaque particule et sa région la plus proche dans le masque d'avant-plan observé. Cependant, afin d'éviter des résultats de suivi erronés, en particulier lorsque aucune région d'avant-plan n'est disponible (par exemple, lors d'une occultation), la technique de fenêtrage (gating) est employée. Celle-ci consiste, dans un premier temps, à construire une fenêtre de recherche autour de la dernière position estimée de l'humain suivi. Ensuite, toutes les régions d'avant-plan qui tombent en dehors de cette fenêtre sont supposées provenir d'autres objets en mouvement plutôt que de l'humain actuellement suivi, et elles sont exclues de l'analyse. Parmi les régions d'avant-plan restantes (c.-à-d., celles qui tombent à l'intérieur de la fenêtre), nous choisissons celle qui est la plus proche en termes de la distance Euclidienne $d_{k,Spatial}^{(i)}$, définie dans l'équation 3.10, de chaque particule du filtre pour évaluer son poids.

$$d_{k,Spatial}^{(i)} = \sqrt{(x_k^{(i)} - x_k^{(i,NNFR)})^2 + (y_k^{(i)} - y_k^{(i,NNFR)})^2} \tag{3.10}$$

Dans l'équation ci-dessus, les vecteurs $(x_k^{(i)}, y_k^{(i)})$ et $(x_k^{(i,NNFR)}, y_k^{(i,NNFR)})$ sont, respectivement, les coordonnées 2D de la i-ème particule et de sa région d'avant-plan la plus proche (*Nearest Neighbor Foreground Region*, *NNFR*).

3.4.3.2. Intensité

L'intensité est la caractéristique la plus fréquemment utilisée pour le suivi visuel d'objets en mouvement en raison de sa robustesse relativement élevée face aux déformations, aux occultations partielles, aux rotations et aux changements d'échelle.

Soit $\hat{h}_{k,Int} = \{\hat{h}_{k,Int}(u)\}_{u=1}^{m1}$ l'histogramme d'intensité, à l'instant k, du modèle de référence de l'humain détecté et qui est automatiquement initialisé pendant l'étape d'initialisation du suivi (Section 3.3.1), et soit $h_{k,Int}^{(i,NNFR)} = \left\{h_{k,Int}^{(i,NNFR)}(u)\right\}_{u=1}^{m1}$ l'histogramme d'intensité, à l'instant k, de la région d'avant-plan la plus proche associée à la i-ème particule à pondérer. La mesure de similarité entre $\hat{h}_{k,Int}$ et $h_{k,Int}^{(i,NNFR)}$ peut être calculée en utilisant la distance de Bhattacharyya définie dans l'équation 3.11 ci-dessous :

$$d_{k,Int}^{(i)}\left(\hat{h}_{k,Int}, \quad h_{k,Int}^{(i,NNFR)}\right) = \sqrt{1 - \sum_{u=1}^{m_1} \sqrt{\hat{h}_{k,Int}(u)\, h_{k,Int}^{(i,NNFR)}(u)}} \qquad (3.11)$$

où, m_1 est le nombre de *bins* dans chacun des histogrammes d'intensité $\hat{h}_{k,Int}$ et $h_{k,Int}^{(i,NNFR)}$.

D'après l'équation 3.11, nous pouvons observer que, plus $\hat{h}_{k,Int}$ et $h_{k,Int}^{(i,NNFR)}$ sont similaires, plus la distance $d_{k,Int}^{(i)}$ est proche de 0.

3.4.3.3. Texture

La texture est une caractéristique visuelle très importante pour la description de l'apparence des objets en mouvement qui présentent un haut degré de similarité dans leurs modèles d'intensité. Dans ce travail, afin d'extraire les caractéristiques de texture des humains à suivre, nous utilisons l'opérateur Rotated Local Binary Pattern (RLBP) proposé par (Mehta and Egiazarian, 2016), qui est une amélioration de l'opérateur original Local Binary Pattern (LBP) introduit par (Ojala et al., 1996).

L'opérateur LBP d'un pixel central d'une image est calculé par le seuillage des différences entre sa valeur d'intensité et les valeurs d'intensité de ses pixels voisins. Les valeurs résultantes sont ensuite considérées comme un "motif binaire",

également connu sous le nom de "code LBP", qui décrit la texture locale de ce pixel. La définition générale de l'opérateur LBP original est donnée par l'équation 3.12 :

$$\mathrm{LBP}_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(\mathrm{g}_p - \mathrm{g}_c)2^p\,, \qquad s(\mathrm{g}_p - \mathrm{g}_c) = \begin{cases} 1 & \text{if } \mathrm{g}_p \geq \mathrm{g}_c \\ 0 & autrement \end{cases} \qquad (3.12)$$

où g_c est la valeur de l'intensité du pixel central (x_c, y_c) d'un voisinage local, et $\{\mathrm{g}_p\}_{p=0}^{P-1}$ sont les valeurs d'intensité des P pixels voisins équidistants sur un voisinage circulaire de rayon R.

La Figure 3.1 présente un exemple de calcul de l'opérateur $\mathrm{LBP}_{8,1}$ (P=8, R=1) pour une petite portion d'une image en niveaux de gris.

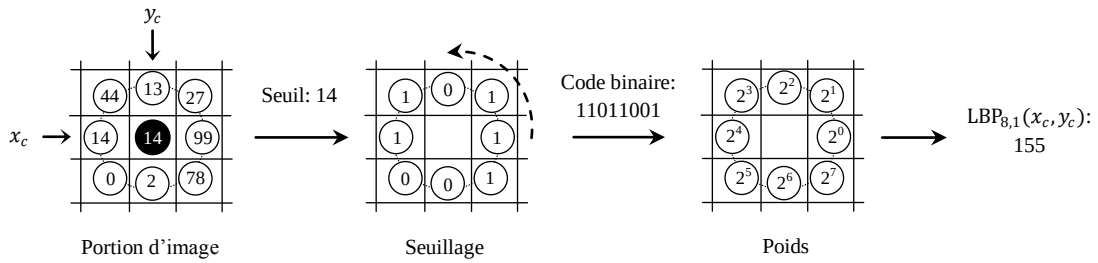


Figure 3.1: Calcul de l'opérateur $\mathrm{LBP}_{8,1}$ pour une petite portion d'une image en niveaux de gris.

L'opérateur $\mathrm{LBP}_{P,R}$ original est relativement insensible aux variations d'illumination, mais il ne possède pas la propriété d'invariance à la rotation. Afin de résoudre ce problème, l'opérateur $\mathrm{RLBP}_{P,R}$ (Rotated Local Binary Pattern) a été proposé dans (Mehta and Egiazarian, 2016). Cet opérateur est défini par l'équation suivante :

$$\mathrm{RLBP}_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(\mathrm{g}_p - \mathrm{g}_c)2^{mod\,(p-D,P)} \qquad (3.13)$$

où, *mod* est l'opérateur *modulo*, et *D*, appelée "direction dominante", est l'indice du pixel voisin pour lequel la différence avec le pixel central est maximale, c'est-à-dire :

$$D = \underset{p\in(0,1,\ldots,P-1)}{\arg\max} \left|\mathrm{g}_p - \mathrm{g}_c\right| \qquad (3.14)$$

Sur la base de l'opérateur RLBP défini dans l'équation 3.13, nous définissons par $\hat{h}_{k,\mathrm{RLBP}} = \{\hat{h}_{k,\mathrm{RLBP}}(u)\}_{u=1}^{m2}$ l'histogramme de texture du modèle de référence à l'instant k de l'humain détecté, et qui est automatiquement initialisé pendant l'étape d'initialisation du suivi (Section 3.3.1), et par $h_{k,RLBP}^{(i,NNFR)} = \left\{h_{k,RLBP}^{(i,NNFR)}(u)\right\}_{u=1}^{m2}$ l'histogramme de texture à l'instant k de la région d'avant-plan la plus proche associée à la *i*-ème particule à pondérer. Ensuite, la mesure de similarité entre $\hat{h}_{k,\mathrm{RLBP}}$ et $h_{k,RLBP}^{(i,NNFR)}$ peut être calculée en utilisant la distance de Bhattacharyya définie dans l'équation 3.15 :

$$d_{k,\mathrm{RLBP}}^{(i)}\left(\hat{h}_{k,RLBP}, h_{k,RLBP}^{(i,NNFR)}\right) = \sqrt{1 - \sum_{u=1}^{m_2}\sqrt{\hat{h}_{k,RLBP}(u)\, h_{k,RLBP}^{(i,NNFR)}(u)}} \qquad (3.15)$$

où m_2 est le nombre de *bins* dans chacun des histogrammes $\hat{h}_{k,RLBP}$ et $h_{k,RLBP}^{(i,NNFR)}$.

Comme la distance définie dans l'équation 3.11, plus les histogrammes $\hat{h}_{k,RLBP}$ et $h_{k,RLBP}^{(i,NNFR)}$ sont similaires, plus la distance $d_{k,\mathrm{RLBP}}^{(i)}$ est proche de 0.

3.4.3.4. Mouvement

La caractéristique de mouvement est obtenue en utilisant la vélocité moyenne de l'humain suivi sur les dernières trames de la séquence d'images. Cette vélocité moyenne est calculée à l'aide de l'équation 3.16 ci-dessous :

$$\overline{\boldsymbol{v}}_{k-1} = \left(\frac{1}{k_0}\sum_{t=k-k_0}^{k-1}\hat{x}_t - \hat{x}_{t-1}, \frac{1}{k_0}\sum_{t=k-k_0}^{k-1}\hat{y}_t - \hat{y}_{t-1}\right)^T \qquad (3.16)$$

où, $(\hat{x}_t, \hat{y}_t)$ est la position estimée, à l'instant *t*, de l'humain suivi, et k_0 est la longueur de l'historique récent de trames considéré pour le calcul de la vélocité moyenne.

Ensuite, soit $v_k^{(i,NNFR)}$ la vélocité de mouvement de la région d'avant-plan la plus proche associée à la *i*-ème particule à pondérer. Cette vélocité est calculée par rapport à la dernière position estimée de l'humain suivi, comme montré dans l'équation 3.17 :

$$\boldsymbol{v}_k^{(i,NNFR)} = \left(x_k^{(i,NNFR)} - \hat{x}_{k-1},\ y_k^{(i,NNFR)} - \hat{y}_{k-1}\right)^T \qquad (3.17)$$

Ainsi, la mesure de similarité entre $\overline{\boldsymbol{v}}_{k-1}$ et $\boldsymbol{v}_k^{(i,NNFR)}$ est obtenue en utilisant la somme pondérée donnée par l'expression suivante :

$$d_{k,Mot}^{(i)}\left(\overline{\boldsymbol{v}}_{k-1}, \boldsymbol{v}_k^{(i,NNFR)}\right) = \xi\, d_{k,Dir}^{(i)} + (1-\xi)\, d_{k,Mag}^{(i)} \tag{3.18}$$

où, les termes $d_{k,Dir}^{(i)}$ et $d_{k,Mag}^{(i)}$, définis respectivement dans les équations 3.19 et 3.20, sont les similarités en direction et en amplitude entre $\overline{\boldsymbol{v}}_{k-1}$ et $\boldsymbol{v}_k^{(i,NNFR)}$, et le paramètre ξ est un facteur de pondération (qui est fixé à 0.5 dans nos expériences).

$$d_{k,Dir}^{(i)}\left(\overline{\boldsymbol{v}}_{k-1}, \boldsymbol{v}_k^{(i,NNFR)}\right) = \frac{1}{\pi} cos^{-1}\left(\frac{\overline{\boldsymbol{v}}_{k-1} \circ \boldsymbol{v}_k^{(i,NNFR)}}{\|\overline{\boldsymbol{v}}_{k-1}\| \left\|\boldsymbol{v}_k^{(i,NNFR)}\right\|}\right) \tag{3.19}$$

$$d_{k,Mag}^{(i)}\left(\overline{\boldsymbol{v}}_{k-1}, \boldsymbol{v}_k^{(i,NNFR)}\right) = 1 - \frac{min\left(\|\overline{\boldsymbol{v}}_{k-1}\|,\ \left\|\boldsymbol{v}_k^{(i,NNFR)}\right\|\right)}{max\left(\|\overline{\boldsymbol{v}}_{k-1}\|,\ \left\|\boldsymbol{v}_k^{(i,NNFR)}\right\|\right)} \tag{3.20}$$

Dans les deux équations ci-dessus, le symbole ∘ représente le produit scalaire de deux vecteurs, et le symbole $\|\cdot\|$ représente la norme Euclidienne d'un vecteur.

3.4.3.5. Combinaison adaptative des caractéristiques

Après avoir calculé les distances $d_{k,Spatial}^{(i)}$, $d_{k,Int}^{(i)}$, $d_{k,RLBP}^{(i)}$ et $d_{k,Mot}^{(i)}$, les poids attribués aux particules du filtre sont estimés à l'aide de la formule gaussienne présentée dans l'équation 3.21 :

$$\boldsymbol{w}_k^{(i)} = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2}\left(\frac{\left(\widehat{\omega}_{k,1}\, d_{k,Int}^{(i)} + \widehat{\omega}_{k,2}\, d_{k,RLBP}^{(i)} + \widehat{\omega}_{k,3}\, d_{k,Mot}^{(i)}\right) \cdot d_{k,Spatial}^{(i)}}{\sigma}\right)^2\right)} \tag{3.21}$$

où, σ dénote l'écart type des erreurs de mesure, et $\widehat{\omega}_{k,1}, \widehat{\omega}_{k,2}$ et $\widehat{\omega}_{k,3}$ sujets à $\widehat{\omega}_{k,1} + \widehat{\omega}_{k,2} + \widehat{\omega}_{k,3} = 1$, sont, respectivement, les poids qui contrôlent la contribution des caractéristiques d'intensité, de texture RLBP et de vélocité de mouvement dans le calcul des poids des particules. Le choix le plus simple que nous pouvons faire dans l'équation 3.21 est de considérer des valeurs constantes pour les poids des caractéristiques $\widehat{\omega}_{k,1}, \widehat{\omega}_{k,2}$ et $\widehat{\omega}_{k,3}$. Cependant, dans des environnements réels non contrôlés, une telle hypothèse n'est pas appropriée puisque la fiabilité, ou la capacité discriminante, de chacune des caractéristiques individuelles, peut changer considérablement pendant le processus de suivi, et ce, en raison de nombreux

facteurs tels que le bruit, les occultations, les changements d'illumination, etc. Ainsi, dans notre approche, au lieu de supposer des valeurs constantes, nous proposons de calculer les poids $\widehat{\omega}_{k,1}, \widehat{\omega}_{k,2}$ et $\widehat{\omega}_{k,3}$ de manière adaptative afin de combiner de manière optimale les avantages des différents types de caractéristiques dans un cadre unifié.

Cependant, dans la pratique, lorsque la fiabilité, à l'instant k-1, d'un type de caractéristique donné est élevée, la mesure de similarité basée sur cette caractéristique doit être pondérée par un facteur de pondération élevé à l'instant k. Inversement, lorsqu'un type de caractéristique est moins fiable à l'instant k-1, la mesure de similarité basée sur cette caractéristique à l'instant k doit être pondérée par un facteur de pondération faible. Ainsi, afin d'évaluer la fiabilité des différentes caractéristiques, nous proposons de calculer d'abord les poids des particules dans chaque caractéristique, séparément, en utilisant la formule de l'équation 3.22 :

$$\omega_{k-1,c}^{(i)} = \frac{1}{\sqrt{2\pi}\sigma_c} e^{\left(-\frac{1}{2}\left(\frac{d_{k-1,c}^{(i)} \cdot d_{k-1,Spatial}^{(i)}}{\sigma_c} \right)^2 \right)} \tag{3.22}$$

où $c = 1,2,3$ correspondent, respectivement, à la caractéristique d'intensité, de texture RLBP et de vélocité du mouvement, le terme $d_{k-1,c}^{(i)}$ est la distance de similarité, à l'instant k-1, basée sur la c-ième caractéristique, et σ_c désigne l'écart type des erreurs de mesure dans la c-ième caractéristique.

Ensuite, dans chacune des caractéristiques individuelles, nous classons toutes les particules dans un ordre décroissant en fonction de leur poids, et la moyenne des poids des N_p premières particules, calculée selon l'équation 3.23, est utilisée comme une estimation de la fiabilité de la caractéristique correspondante.

$$\overline{\omega}_{k,c} = \frac{1}{N_p} \sum_{n=1}^{N_p} \omega_{k-1,c}^{(n)}, \quad c = 1,2,3 \tag{3.23}$$

Pour satisfaire la contrainte $\sum_{c=1}^{3} \omega_{k,c} = 1$, les poids des différentes caractéristiques sont normalisés comme suit:

$$\widehat{\omega}_{k,c} = \frac{\overline{\omega}_{k,c}}{\sum_{c=1}^{3} \overline{\omega}_{k,c}}, \qquad c = 1,2,3 \tag{3.24}$$

D'après l'équation 3.23, nous pouvons observer que plus la fiabilité, ou la capacité discriminante d'une caractéristique donnée à l'instant précédent est élevée, plus les poids des particules basés sur cette caractéristique sont élevés et, par conséquent, plus la valeur de confiance attribuée à cette caractéristique à l'instant courant est importante. Inversement, plus la fiabilité d'une caractéristique à l'instant précédent est faible, moins la valeur de confiance qui lui est attribuée à l'instant présent est importante.

3.4.4. Estimation d'état

Après avoir calculé les poids des particules à l'aide de la formule de l'équation 3.21, la position courante $\hat{\mathbf{x}}_k$ de l'humain suivi est estimée selon l'équation 3.25, en multipliant toutes les positions des particules par leurs poids respectifs.

$$\hat{\mathbf{x}}_k = \sum_{i=1}^{N_s} \widehat{\boldsymbol{w}}_k^{(i)} \mathbf{x}_k^{(i)} \tag{3.25}$$

Avec, $\widehat{\boldsymbol{w}}_k^{(i)} = \boldsymbol{w}_k^{(i)} / \sum_{i=1}^{N_s} \boldsymbol{w}_k^{(i)}$ sont les poids normalisés des particules à l'instant k.

3.4.5. Ré-échantillonnage des particules

Le ré-échantillonnage est une étape très importante dans le filtrage à particules, car il permet à l'algorithme d'éviter le phénomène de dégénérescence, c'est-à-dire la convergence de tous les poids des particules, sauf un, vers zéro. La technique de ré-échantillonnage la plus couramment utilisée est la méthode de ré-échantillonnage par importance séquentielle (Kuptametee and Aunsri, 2022). Cette méthode est décrite comme suit.

Tout d'abord, la fonction de distribution cumulative (Cumulative Distribution Function, CDF) des poids normalisés des particules est calculée comme montré dans l'équation 3.26 :

$$\mathrm{CDF}_k^{(i)} = \sum_{h=1}^{i} \widehat{\boldsymbol{w}}_k^{(h)}, \quad i = 1, \dots, N_s \tag{3.26}$$

Ensuite, un nombre aléatoire $\vartheta_{j \in [1,\dots,N_s]}$, est tiré aléatoirement avec une densité uniforme sur l'intervalle [0,1], puis il est projeté vers le co-domaine de la CDF, puis vers le domaine de la CDF. L'intersection avec le domaine de la CDF constitue un

indice i, et la particule correspondante $\mathbf{x}^{*}{}_{k}^{(i)}$ est sélectionnée et ajoutée au nouvel ensemble (ré-échantillonné) de particules. Une illustration de la procédure de ré-échantillonnage est donnée dans la Figure 3.2.

Nous pouvons constater, d'après le processus de ré-échantillonnage, que les particules ayant un poids élevé seront sélectionnées plusieurs fois, conduisant à des copies identiques, tandis que celles ayant un poids négligeable seront choisies peu de fois, voire jamais. À la fin de la procédure de ré-échantillonnage, l'ancien ensemble de particules est remplacé par le nouveau, et tous les poids des particules sont mis égaux à $1/N_s$.

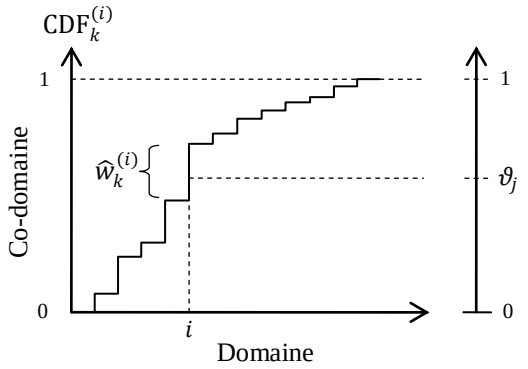


Figure 3.2: Illustration de la procédure de ré-échantillonnage.

3.4.6. Mise à jour du modèle

Dans des scénarios pratiques et réels de vidéo surveillance, l'apparence visuelle de l'humain suivi peut facilement subir des variations dues à de nombreux facteurs, tels que la non-rigidité du corps humain, les occultations, les changements de point de vue, le bruit et les changements d'illumination. Ainsi, une stratégie de mise à jour du modèle humain est nécessaire afin d'éviter l'introduction dans ce modèle d'éléments susceptibles de provoquer des erreurs durant le processus de suivi. Pour atteindre cet objectif, nous proposons dans cette section une stratégie automatique qui effectue l'opération de mise à jour progressivement en fonction de la distance de similarité entre l'apparence actuelle de l'humain suivi et le modèle de référence. Cette stratégie est décrite comme suit.

Tout d'abord, supposant que $h_{k,Int}^{Est}$ et$h_{k,RLBP}^{Est}$ sont, respectivement, les histogrammes d'intensité et de texture RLBP de l'humain suivi, et qui sont obtenus sur la base de la dernière position estimée $\hat{\mathrm{x}}_k$ en utilisant un filtre à particules. Ensuite, les

histogrammes (modèles) de référence d'intensité $\hat{h}_{k,Int}$ et de texture RLBP $\hat{h}_{k,RLBP}$ sont, respectivement, mis à jour à l'aide des filtres récursifs à réponse impulsionnelle infinie définis dans les équations 3.27 et 3.28 :

$$\hat{h}_{k+1,Int} = \left(1 - \alpha_{k,Int}\right) \hat{h}_{k,Int} + \alpha_{k,Int}\ h_{k,Int}^{Est} \tag{3.27}$$

$$\hat{h}_{k+1,RLBP} = \left(1 - \alpha_{k,RLBP}\right) \hat{h}_{k,RLBP} + \alpha_{k,RLBP}\ h_{k,RLBP}^{Est} \tag{3.28}$$

où, $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ sont les taux qui contrôlent la vitesse de mise à jour des modèles.

La sélection de valeurs appropriées pour les deux paramètres $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ est très importante, car si nous choisissons une valeur trop élevée, le modèle sera surajusté et cela peut éventuellement conduire à une mauvaise performance de suivi au cours du temps. D'autre part, si nous choisissons une valeur trop faible, le modèle sera incapable de capturer les variations importantes et rapides de l'apparence humaine, et cela peut entraîner une dégradation significative des performances, voire l'échec de l'ensemble du processus de suivi. Ainsi, pour faire face à ce problème, nous proposons d'ajuster les valeurs de $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ automatiquement, en fonction de la mesure de similarité entre l'apparence actuelle de l'humain suivi et le modèle de référence. La procédure à suivre pour y parvenir est la suivante. Supposons, tout d'abord, que $d_{k,Int}\left(h_{k,Int}^{Est},\ \hat{h}_{k,Int}\right)$ et $d_{k,RLBP}\left(h_{k,RLBP}^{Est},\ \hat{h}_{k,RLBP}\right)$ sont, respectivement, les distances de Bhattacharyya, basées sur les caractéristiques d'intensité et de texture RLBP, entre l'histogramme actuel de l'humain suivi et l'histogramme (modèle) de référence. Ensuite, les valeurs des taux de mise à jour $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ sont, respectivement, calculées en utilisant les fonctions exponentielle décroissante suivantes :

$$\alpha_{k,Int} = \alpha_{Max}\ e^{\left(-\gamma\ d_{k,Int}\left(h_{k,Int}^{Est},\hat{h}_{k,Int}\right)\right)} \tag{3.29}$$

$$\alpha_{k,RLBP} = \alpha_{Max}\ e^{\left(-\gamma\ d_{k,RLBP}\left(h_{k,RLBP}^{Est},\hat{h}_{k,RLBP}\right)\right)} \tag{3.30}$$

où, γ est un paramètre positif qui contrôle la vitesse de décroissance, et α_{Max} est la valeur maximale que les taux de mise à jour $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ peuvent atteindre, et son rôle est d'empêcher le modèle de référence d'être excessivement mis à jour. Dans nos expériences, γ et α_{Max} sont fixés à 6 et 0,1, respectivement.

Il est essentiel de remarquer d'après les équations 3.29 et 3.30 que, lorsque la similarité entre l'histogramme d'intensité (resp. de texture RLBP) de l'humain actuellement suivi et l'histogramme d'intensité (resp. de texture RLBP) du modèle de référence est élevée, le modèle sera mis à jour rapidement avec une valeur élevée du taux de mise à jour. Inversement, lorsque la similarité est faible (par exemple, lors d'un changement important d'illumination), le modèle de référence sera mis à jour lentement avec une petite valeur du taux de mise à jour.

3.4.7. Détection et gestion des occultations

Les occultations sont un problème sérieux et très courant, qui est difficile à éviter dans le cadre du suivi visuel, notamment lorsqu'il s'agit de plusieurs personnes en mouvement dans la scène surveillée. Typiquement, dans les scénarios réels de vidéo surveillance, nous pouvons distinguer deux types d'occultations : l'occultation inter-humain et l'occultation humain/arrière-plan. Le premier type, l'occultation inter-humain, se produit lorsque plusieurs personnes suivies indépendamment se croisent, tandis que le second type, l'occultation humain-arrière-plan, se produit lorsque certains objets statiques de l'arrière-plan occultent l'humain suivi. Ainsi, si nous excluons ces deux problèmes sérieux de l'analyse, les performances du processus de suivi peuvent diminuer considérablement avec le temps et peuvent même entraîner la perte totale de l'humain suivi. C'est pourquoi il est nécessaire de mettre en place une procédure de détection et de traitement des occultations humaines afin d'augmenter la robustesse de l'algorithme de suivi lorsqu'une situation d'occultation se produit.

3.4.7.1. Occultation inter-humain

Dans notre système proposé, une occultation inter-humain est détectée lorsque deux humains actuellement suivis satisfont les deux conditions suivantes.
La première condition est basée sur l'application d'un seuil sur la distance Euclidienne entre les positions estimées de deux humains actuellement suivis. En général, lorsqu'une occultation inter-humain se produit, les positions estimées des humains entrés en occultation ont tendance à être très proches les unes des autres. Cette situation peut donc être détectée à l'aide de la formule de l'équation 3.31 :

$$\sqrt{\left(\hat{x}_k^i - \hat{x}_k^j\right)^2 + \left(\hat{y}_k^i - \hat{y}_k^j\right)^2} \leq d_{Th} \qquad (3.31)$$

où, $(\hat{x}_k^i, \hat{y}_k^i)$ et $(\hat{x}_k^j, \hat{y}_k^j)$ sont respectivement les positions estimées, à l'instant k, du i-ième et du j-ième humain actuellement suivis, et d_{Th} est un seuil de distance déterminé expérimentalement.

La deuxième condition est basée sur la mesure de la variation des valeurs de la hauteur et de la largeur des humains pendant le processus de suivi. En général, lorsqu'une occultation inter-humain se produit, la largeur, ou éventuellement la hauteur, des humains entrés en occultation varient largement en raison de la fusion de leurs régions d'avant-plan (silhouettes) correspondantes. Cette situation peut donc être détectée à l'aide de la formule de l'équation 3.32 :

$$\left[\left|H_k^i - \bar{H}_k^i\right| \geq 2\sigma_{H_k^i} \wedge \left|H_k^j - \bar{H}_k^j\right| \geq 2\sigma_{H_k^j}\right] \vee \left[\left|W_k^i - \bar{W}_k^i\right| \geq 2\sigma_{W_k^i} \wedge \left|W_k^j - \bar{W}_k^j\right| \geq 2\sigma_{W_k^j}\right] \quad (3.32)$$

où, (H_k^i, W_k^i) et (H_k^j, W_k^j) sont, respectivement, les valeurs de la hauteur et de la largeur à l'instant k du i-ème et du j-ème humain suivi, $(\bar{H}_k^i, \bar{W}_k^i, \sigma_{H_k^i}, \sigma_{W_k^i})$ et $(\bar{H}_k^j, \bar{W}_k^j, \sigma_{H_k^j}, \sigma_{W_k^j})$ sont, respectivement, leur hauteur moyenne, leur largeur moyenne, leur écart-type de hauteur et leur écart-type de largeur sur les trames récentes de la séquence d'images IR.

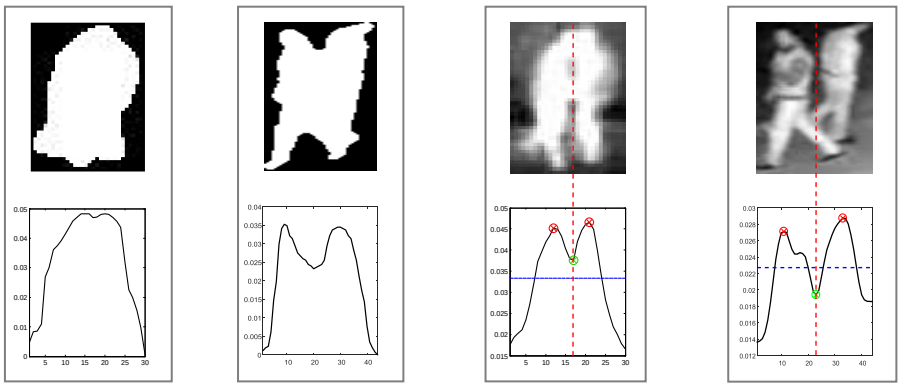
Ainsi, lorsque les deux conditions ci-dessus sont simultanément satisfaites pour deux humains (ou plus) actuellement suivis, une occultation inter-humain est détectée. Dans ce cas, une procédure de séparation des humains entrés en occultation est exécutée. Actuellement, la technique la plus couramment utilisée pour cette tâche est celle basée sur l'Histogramme de Projection Verticale (HPV) de la région binaire d'avant-plan (silhouettes) des humains entrés en occlusion (J. Wang et al., 2012; Jeon et al., 2016). Cependant, dans certains cas, notamment lorsque l'image d'avant-plan extraite par la technique de soustraction d'arrière-plan n'est pas suffisamment précise, la séparation basée sur le HPV devient difficile à réaliser. D'autre part, l'une des caractéristiques principales du corps humain dans les images IR est sa forte intensité due à sa forte émissivité (en cas d'images thermiques), ou réflectivité (en cas d'images proche IR) de rayonnements IR. Dans ce cas, lorsqu'une occlusion inter-humain se produit, l'espace (ou l'écart) entre les humains entrés en occultation apparaît plus clairement. Ainsi, basé sur cette

propriété, et afin de séparer avec précision les humains entrés en occultation, nous proposons dans ce travail de calculer le HPV directement à partir de l'image en niveaux de gris contenant les humains entrés en occultation, plutôt qu'à partir de l'image binaire d'avant-plan produite par la technique de soustraction d'arrière-plan. Les étapes de la procédure de séparation sont les suivantes :

1) Calculer le HPV en niveaux de gris par une somme en colonnes des valeurs des pixels à l'intérieur de la boîte englobante minimale contenant les humains entrés en occultation.
2) Trouver tous les pics de HPV en niveaux de gris qui dépassent un seuil spécifié (ici, la moyenne de HPV en niveaux de gris).
3) Trouver tous les points de vallée significatifs entre les points de pic détectés.
4) Séparer les humains entrés en occultation par des lignes verticales aux positions des points de vallée trouvés.

La Figure 3.3 illustre la procédure de séparation décrite ci-dessus lorsqu'elle est appliquée à des exemples d'humains entrés en occultation. Les figures 3.3 (a–b) représentent les masques d'avant-plan imprécis d'humains avec leur HPV binaires correspondants, alors que les figures 3.3 (c–d) représentent des images thermiques et proche IR, respectivement, avec leur HPV en niveaux de gris correspondants. Sur ces figures, nous pouvons observer que les HPV en niveaux de gris présentent deux pics majeurs et une vallée. Les pics, indiqués par les croix rouges encerclées, correspondent aux deux humains entrés en occultation, tandis que la vallée, indiquée par la croix verte encerclée, correspond à l'espace ou le vide entre eux. La ligne pointillée bleue horizontale indique la valeur moyenne des HPV en niveaux de gris, et la ligne pointillée rouge verticale indique l'emplacement de la séparation.

Cependant, lorsque la procédure de séparation ne parvient pas à trouver des points de pic ou de vallée significatifs dans le HPV en niveaux de gris, nous considérons que les humains sont fortement occultés. Dans ce cas, la mise à jour des poids des particules est effectuée en utilisant uniquement les caractéristiques de proximité spatiale et de mouvement, c'est-à-dire que les poids $\widehat{\omega}_{k,1}$ et $\widehat{\omega}_{k,2}$ dans l'équation 3.21 sont mis à zéro.



(a) (b) (c) (d)

Figure 3.3 : Procédure de séparation d'humains entrés en occultation. (a) et (b) Masques d'avant-plan imprécis avec leur HPV binaires correspondants. (c) et (d) Images thermique et proche IR, respectivement, avec leur HPV en niveaux de gris correspondants.

En outre, afin d'éviter l'introduction des erreurs dans le modèle de référence, la procédure de mise à jour des histogrammes d'intensité $\hat{h}_{k+1,Int}$ et de texture $\hat{h}_{k+1,RLBP}$ est désactivée, et ce, en mettant à zéro les valeurs des paramètres $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ dans les équations 3.27 et 4.28. Une fois l'occultation inter-humain est terminée, la procédure de mise à jour repasse en mode normal.

3.4.7.2. Occultation humain/arrière-plan

Une occultation humain/arrière-plan est détectée lorsqu'aucune région d'avant-plan n'est trouvée dans la fenêtre de recherche autour de la dernière position estimée de l'humain suivi. Dans ce cas, l'estimation de la position de l'humain est effectuée en utilisant uniquement l'étape de prédiction (sous-section 3.3.2), tandis que l'étape de mise à jour des poids des particules (sous-section 3.3.3) est temporairement suspendue jusqu'à ce que l'occultation humain/arrière-plan disparaisse. Cependant, si l'occultation humain/arrière-plan détectée persiste sur une durée suffisamment longue, ou si la position de l'humain est prédite en dehors du plan image, nous considérons que l'humain suivi a probablement quitté le champ de vision de la caméra. Dans ce cas, sa piste (track) est supprimée de la liste des pistes actuelles et le filtre à particules correspondant est ré-initialisé pour suivre un nouvel humain entrant dans la scène dans les trames suivantes.

3.5. Conclusion

Dans ce chapitre, nous avons tout d'abord présenté les différentes méthodes qui ont été proposées dans la littérature pour le suivi de personnes dans des séquences d'images IR. Nous avons ensuite présenté notre nouvelle approche qui est basée sur le filtre à particules et une combinaison d'informations provenant de plusieurs types de caractéristiques, à savoir l'intensité, la texture, la vélocité de mouvement, et la distance spatiale. La combinaison de ces caractéristiques est effectuée de manière adaptative en tenant compte de la capacité discriminative de chaque caractéristique au cours du processus de suivi. Afin de rendre la méthode proposée plus robuste notamment face aux déformations et aux changements d'apparence, nous avons introduit une procédure de mise à jour automatique du modèle des personnes suivies dans la scène surveillée. Dans ce chapitre, nous avons également présenté une nouvelle stratégie pour détecter et gérer les occlusions en utilisant des règles heuristiques simples et l'histogramme de projection verticale (VPH) en niveaux de gris. Dans le prochain chapitre, nous allons présenter la troisième partie de notre système proposé, à savoir la reconnaissance de posture humaine.

Chapitre 4

Reconnaissance de posture humaine

4.1. Introduction

La posture humaine est définie comme étant la façon dont une personne détient son corps (Lim et al., 2018). La reconnaissance automatique de posture humaine est un sujet de recherche important qui trouve des applications pratiques dans de nombreux domaines tels que la vidéo surveillance, l'interaction homme-robot, la recherche d'images et de vidéos par le contenu, le sport et exercice physique, les véhicules intelligents, l'analyse de l'activité et du comportement humain, les soins de santé, etc. Dans ce contexte d'application, les approches proposées peuvent être regroupées en deux grandes catégories. La première catégorie, appelée approches basées sur des capteurs (sensor-based approaches), repose sur l'utilisation d'un ensemble de capteurs de mouvement portables, tels que des accéléromètres (J. Wang et al., 2016), des gyroscopes (Idris et al., 2015), ou leurs combinaisons (Tanaka et al., 2004) pour détecter et suivre le mouvement, et localiser les différentes articulations ou parties du corps humain. Ces approches donnent des performances satisfaisantes, mais leur principal inconvénient est qu'elles nécessitent la fixation de capteurs sur le corps humain, qui sont souvent gênants à porter et exigent des batteries qui doivent être rechargées ou remplacées périodiquement pour une performance optimale. La deuxième catégorie, appelée

approches basées sur la vision (vision-based approaches), repose sur l'utilisation d'une ou plusieurs caméras pour capturer des images statiques ou des séquences d'images du corps humain. Ces images ou séquences d'images sont ensuite analysées à l'aide de techniques de vision par ordinateur pour estimer la posture humaine. Les approches appartenant à cette catégorie peuvent être encore subdivisées en deux groupes principaux, à savoir les approches basées sur des marqueurs (marker-based approaches) et les approches sans marqueurs (marker-free approaches). Les approches basées sur des marqueurs (Ukida et al., 2006; Yang et al., 2011) consistent à placer des marqueurs spéciaux sur certains points importants du corps humain, tels que les articulations et la tête, ce qui facilite le processus d'estimation de la posture du corps humain sans la nécessité de traitements intensifs. Ces approches ont démontré leur capacité à reconnaître plusieurs types de postures humaines, mais leur principal inconvénient est qu'elles nécessitent de placer des marqueurs sur le corps humain, ce qui réduit le caractère naturel du processus de reconnaissance de posture. De plus, comme dans les approches basées sur des capteurs portables, ces marqueurs peuvent être très inconfortables pour le sujet humain, et ils nécessitent d'être bien observés par la caméra, ce qui est souvent difficile à atteindre, notamment lorsqu'une partie du corps humain est partiellement occultée par d'autres parties ou par des objets voisins dans la scène. Quant aux approches sans marqueurs, elles n'exigent ni le port de marqueurs, ni de capteurs, sur le corps humain. Elles exploitent uniquement des méthodes de traitement d'image et de vision par ordinateur pour effectuer la tâche de reconnaissance de posture.

De nombreuses approches de reconnaissance de posture humaine sans marqueurs basées sur des caractéristiques et des modèles d'apprentissage différents ont été proposées dans la littérature (Cucchiara et al., 2005; Boulay et al., 2006; Juang and Chang, 2007; Chen et al., 2012; Juang et al., 2014; Wang et al., 2016; Kang and Lee, 2016; Zerrouki and Houacine, 2018;). Dans ce chapitre, nous proposons une nouvelle méthode basée sur un SVM multi-classe et une combinaison de trois caractéristiques différentes, à savoir les moments de Krawtchouk, l'histogramme de chaîne de code et des caractéristiques géométriques. Chaque type de ces caractéristiques décrit différents aspects du contenu de la silhouette humaine. Les moments de Krawtchouk, par exemple, décrivent la région à l'intérieur de la silhouette humaine, alors que l'histogramme de chaîne de code décrit les points du

contour qui l'entoure. Quant aux caractéristiques géométriques, elles décrivent l'information globale du corps humain.
Le reste de ce chapitre est organisé comme suit. Un état de l'art des méthodes proposées dans la littérature pour la reconnaissance de posture humaine sera tout d'abord présenté. Nous détaillons ensuite la nouvelle approche que nous proposons, qui est basée sur une combinaison de trois caractéristiques différentes, à savoir des caractéristiques basées région, des caractéristiques basées contour et des caractéristiques géométriques. Enfin, nous terminerons ce chapitre par une conclusion qui récapitule les principales contributions.

4.2. Etat de l'art sur la reconnaissance de posture humaine

Avec l'augmentation de la puissance de calcul des machines informatiques, plusieurs nouvelles méthodes ont été proposées et continuent d'être le sujet de recherches dans le domaine de vision par ordinateur et la reconnaissance de formes. Ainsi, depuis quelques années, de nombreux chercheurs ont concentré leur attention sur le développement de méthodes efficaces de reconnaissance de postures humaines basées sur la vision, en vue de leur application dans des environnements réels. Le défi majeur est de prendre en considération le nombre et la diversité de postures en raison de la grande flexibilité du corps humain. La présence de bruits dans les images, ainsi que des changements d'illumination et d'environnement sont d'autres défis à relever. Afin de surmonter certains de ces défis, de nombreuses approches de reconnaissance de posture humaine sans marqueurs basées sur des caractéristiques et des modèles d'apprentissage (classifieurs) différents ont été proposées dans la littérature (Ma et al., 2022). Selon la dimension des données vidéo utilisées, ces approches peuvent être regroupées en deux grandes catégories : les approches 2D et les approches 3D.

4.2.1. Approches 2D

Les approches 2D reposent sur l'utilisation des caractéristiques d'apparence 2D du corps humain. Haritaoglu et al., (1998), par exemple, ont proposé d'utiliser les Histogrammes de Projection Horizontale et Verticale (HVPH) et une classification hiérarchique pour déterminer la posture principale de l'humain détecté. En faisant correspondre cette posture à certains points caractéristiques du contour de la silhouette humaine, les parties les plus significatives du corps humain, telles que la

tête, les mains, le torse et les pieds, sont localisées. Cette approche a la capacité de gérer des arrière-plans complexes, mais elle nécessite des améliorations pour le traitement des ombres. Les HVPH sont également utilisés par Goldmann et al., (2004) en combinaison avec le descripteur Curvature Scale Space (CSS). Deux classifieurs, à savoir le classifieur de distance minimale et l'algorithme des *k*-plus proches voisins (*k*-Nearest Neighbor, *k*NN) sont utilisés pour reconnaître, simultanément, la posture principale et le point de vue de l'humain détecté. Afin d'obtenir une plus grande robustesse face aux mouvements non rigides du corps humain, Cucchiara et al., (2005) ont utilisé les histogrammes de projection de la silhouette humaine et un processus d'apprentissage automatique supervisé pour créer des cartes de projection probabilistes (Probabilistic Projection Maps, PPM). Ces cartes sont ensuite utilisées comme des caractéristiques d'entrée pour un classifieur bayésien qui utilise un graphe de transition d'état pour estimer la posture de la personne observée. Cette approche peut gérer de manière fiable les occlusions, mais son principal inconvénient est qu'elle nécessite une segmentation parfaite de la silhouette humaine. Takahashi and Naemura, (2005) ont proposé une approche pour estimer la posture humaine en utilisant un réseau de neurones artificiels (Artificial Neural Network, ANN) et un filtre de Kalman. Les entrées du réseau sont les positions relatives du contour échantillonné par rapport au centroïde de la silhouette, alors que les sorties sont les coordonnées 2D des points les plus significatifs du corps humain, tels que la tête, les épaules, les mains, les coudes, les genoux et les pieds. Le filtre de Kalman est introduit par les auteurs pour optimiser et suivre la position des points significatifs estimés. Buccolieri et al., (2005) ont utilisé des contours actifs de type Gradient Vector Flow (GVF) et un réseau de neurones à base radiale (Radial Basis Function, RBF) pour reconnaître trois postures humaines, à savoir "Standing", "Bending" et "Squatting". Cette approche a l'avantage d'être rapide et moins sensible au bruit, mais elle ne permet que la reconnaissance de trois postures de base. Li and Chen, (2006) ont appliqué des règles simples sur certains paramètres tels que les longueurs et les plus grandes largeurs des parties inférieures et supérieures du corps humain afin de reconnaître quatre postures, à savoir "Standing", "Sitting", "Kneeling" et "Stooping". Boulay et al., (2006) ont proposé d'utiliser plusieurs caractéristiques 2D, à savoir les HVPHs, les moments de Hu et le squelette de la silhouette pour représenter la posture humaine. La reconnaissance de posture est effectuée en comparant la

silhouette humaine observée aux silhouettes 2D obtenues à partir d'une projection, sur un espace 2D, d'un ensemble de postures modèles 3D. Cette approche donne des résultats satisfaisants même en cas de présence d'erreurs de segmentation, mais son inconvénient est qu'elle nécessite un temps de calcul relativement élevé. Pour leur part, Girondel et al., (2005) ont décrit la posture humaine en utilisant trois caractéristiques, à savoir la boîte englobante verticale, la boîte des axes principaux et la localisation du visage. Quatre postures prédéfinies sont reconnues en utilisant un classifieur basé sur la théorie de Dempster-Shafer. Cette approche fournit de bons résultats de reconnaissance, mais son principal défaut est sa sensibilité aux changements de la distance caméra/humain. Juang and Chang, (2007) ont présenté une méthode qui utilise un réseau d'inférences neuro-flou auto-constructeur (Self-constructing Neural Fuzzy Inference Network, SONFIN) pour reconnaître quatre postures humaines principales, à savoir "Standing", "Bending", "Sitting" et "Lying". Les caractéristiques calculées sont le rapport hauteur-largeur et les coefficients de la transformation discrète de Fourier (DFT) des histogrammes de projection de la silhouette. La limite de cette méthode est qu'elle nécessite une segmentation précise de la silhouette humaine, ce qui est difficile à atteindre dans les applications du monde réel. Tahir et al., (2007) ont utilisé deux types de filtres de corrélation avancés, à savoir le filtre MACE (Minimum Average Correlation Energy) et le filtre UMACE (Un-constrained Minimum Average Correlation Energy). Dans cette approche, aucune étape d'extraction de caractéristiques n'est effectuée, car les pixels de l'image sont directement utilisés comme vecteur d'entrée des filtres. Dans leur travail, Singh et al., (2008) ont proposé une approche pour la reconnaissance de l'activité humaine basée sur la directionnalité de la silhouette humaine. Dans cette approche, tout d'abord, le contour de la silhouette humaine est représenté comme une chaîne de codes à partir de laquelle les vecteurs directionnels sont extraits. Ensuite, la distribution distincte de ces vecteurs dans l'espace de données est utilisée pour regrouper et reconnaître la posture humaine. Cette approche permet de gérer les changements d'échelle et d'angle de vue, mais l'ensemble de données d'apprentissage nécessite d'être modifié pour des personnes dont leurs silhouettes sont très différentes. En utilisant des règles heuristiques basées sur des caractéristiques de forme et la couleur de la peau, Juang et al., (2009) ont proposé d'estimer la posture en localisant cinq points significatifs du corps humain, à savoir la tête, les extrémités des deux mains et les extrémités des

deux pieds. L'avantage de cette approche est qu'aucune connaissance a priori n'est requise. Xie et al., (2011) ont proposé d'utiliser trois ensembles différents de caractéristiques, à savoir le rapport des largeurs, le rapport des pics et le rapport des valeurs moyennes des HVPHs. Ces caractéristiques sont introduites comme entrées dans une machine à vecteur de support (SVM) pour reconnaître trois types de postures humaines. Un ensemble de filtres de Gabor est utilisé par Chen et al., (2012) pour extraire des caractéristiques de ligne des images de postures humaines. La reconnaissance de posture est effectuée en calculant une mesure de distance de Hausdorff modifiée entre les segments de ligne de l'image de la posture à reconnaître et ceux d'un ensemble de données d'apprentissage. Cette approche est invariante aux translations et aux changements d'échelle, mais pas aux changements de point de vue. Afin de reconnaître les postures humaines dans le contexte de vidéo surveillance de personnes âgées, Brulin et al., (2012) ont proposé une approche basée sur la théorie des ensembles flous et l'analyse des variations de quelques caractéristiques, telles que la hauteur et la largeur de la boîte englobante minimale, le centre de gravité et les projections orthogonales du premier axe principal de la silhouette humaine. Cette méthode est robuste aux changements de distance entre la caméra et la personne surveillée, mais sa principale limite est sa sensibilité aux ombres et aux variations d'intensité. Yu et al., (2012) ont utilisé comme caractéristiques, la meilleure ellipse ajustée de la silhouette humaine et l'histogramme de projection le long des axes de l'ellipse pour reconnaître quatre postures, à savoir "Standing", "Bending", "Sitting" et "Lying". Dans cette approche, la classification de postures est effectuée à l'aide d'un SVM basé sur le graphe de décision acyclique orienté (Directed Acyclic Graph Support Vector Machine, DAGSVM). Li and Sun, (2013) ont combiné trois ANNs dont leurs entrées sont la silhouette, le squelette et les moments de Hu. Les sorties des ANNs sont fusionnées en utilisant la théorie de Dempster-Shafer pour obtenir le résultat final de classification de posture. Un réseau ANN est également utilisé par Zerrouki and Houacine, (2014), où les caractéristiques extraites sont les coefficients de la décomposition en valeurs singulières (Singular Value Decomposition, SVD). Ces mêmes auteurs, dans (Zerrouki and Houacine, 2018), ont présenté une autre approche pour la reconnaissance de posture humaine en utilisant la transformée en curvelet et les aires d'occupation partielle de la silhouette du corps humain. Ces deux approches ont montré des résultats très satisfaisants, mais elles ne sont pas

suffisamment robustes contre les erreurs de segmentation. Hoa and Bui, (2016) ont utilisé deux caractéristiques simples extraites en comptant le nombre de pixels de la silhouette du corps humain. La classification de posture est réalisée à l'aide d'un réseau de neurones flous. De leur côté, Kang and Lee, (2016) ont utilisé la transformée en cosinus discrète (Discrete Cosine Transform, DCT) pour extraire des caractéristiques représentatives de la posture humaine. Un réseau ANN est ensuite utilisé pour reconnaître cinq types différents de posture humaine. Nadeem et al., (2021) ont proposé une méthode qui utilise des opérations de prétraitement pour extraire les silhouettes humaines. Des modèles de parties du corps sont ensuite utilisés pour extraire douze parties clés du corps humain. Ces parties clés sont optimisées pour aider à la génération de caractéristiques multidimensionnelles incluant l'énergie, le flux optique et des caractéristiques distinctives du mouvement. Ces caractéristiques multidimensionnelles sont enfin introduites dans l'Analyse Discriminante Quadratique (ADQ) afin de reconnaître les postures humaines. Plus récemment, Younsi et al., (2023a) ont évalué et comparé différents types de moments orthogonaux, tels que les moments de Zernike, les moments de Legendre et les moments de Chebyshev, et ils ont montré que les moments de Kawtchouk et les moments de Hahn donnent les meilleurs résultats en termes de précision et robustesse face au bruit et aux erreurs de segmentation.

Dans l'ensemble, les approches 2D présentent l'avantage d'avoir un faible coût de calcul, et elles sont donc bien adaptées aux applications en temps réel. Cependant, ces approches sont fortement dépendantes du point de vue de la personne lorsqu'une seule caméra est utilisée.

4.2.2. Approches 3D

Pour atteindre une reconnaissance de posture plus précise et moins dépendante aux changements de point de vue, plusieurs approches utilisant les informations 3D fournies par des caméras stéréo, des caméras multiples ou des caméras de profondeur telles que la Kinect (Younsi et al., 2023b) ou les caméras à temps de vol (Time-of-Flight, ToF) (Diraco et al., 2013) ont été proposées. Pellegrini and Iocchi, (2008), par exemple, ont proposé une approche qui utilise une caméra stéréo pour le suivi et la reconnaissance de la posture humaine. Dans cette approche, une variante de l'algorithme ICP (Iterative Closest Point) est d'abord appliquée pour faire correspondre les données 3D avec un modèle 3D du corps

humain. Ensuite, afin d'augmenter la robustesse contre le bruit de perception, un suivi 3D de trois points principaux (la tête, le bassin et les jambes) dans le modèle est effectué en utilisant un ensemble de filtres de Kalman. Une méthode de classification basée sur un modèle de Markov caché (Hidden Markov Model, HMM) est enfin employée pour reconnaître la posture de l'humain suivi. Cette approche est assez robuste aux occlusions et aux différents points de vue, mais elle est sensible aux erreurs dues à une mauvaise segmentation. Pour reconnaître la posture 3D d'un humain, Hu et al., (2007) ont utilisé des images 2D collectées à partir de différents angles de vue de la caméra. Les descripteurs de Fourier du contour de la silhouette sont extraits en tant que caractéristiques et la reconnaissance de posture humaine est effectuée en utilisant une approche de graphe d'aspect basée sur la similarité. À partir des contours 2D de la silhouette humaine pris avec deux caméras, Juang et al., (2014) ont proposé d'estimer la posture 3D en localisant plusieurs points significatifs, tels que la tête, le centre du corps, le bout des mains, le bout des pieds, les coudes et les genoux. Ensuite, une approche d'appariement basée sur un filtre de Kalman est introduite pour reconstruire les emplacements des points significatifs du corps humain dans l'espace 3D en utilisant les résultats d'estimations de posture en 2D. Cette approche s'est avérée efficace pour l'estimation de la posture humaine en 3D. Cependant, la procédure d'extraction des points significatifs peut échouer pour certaines postures non générales. Dans leur approche basée sur une caméra à temps de vol (ToF), Wientapper et al., (2009) ont appliqué de multiples étapes de prétraitement aux données ToF pour obtenir des images de caractéristiques de basse résolution. Ensuite, pour reconnaître la posture humaine, les auteurs ont utilisé trois techniques de projection linéaire, à savoir l'analyse en composantes principales (Principal Component Analysis, PCA), l'analyse discriminante linéaire (Linear Discriminant Analysis, LDA) et les projections préservant la localité (Locality Preserving Projections, LPP). Le et al., (2013) ont proposé une approche pour la reconnaissance de posture humaine en 3D en utilisant les différents angles entre les joints du squelette humain fournis par la caméra Kinect. Un classifieur SVM multi-classe est utilisé pour reconnaître quatre postures humaines de base. Une autre approche basée sur les images de profondeur capturées par la caméra Kinect est proposée par Wang et al., (2016). Dans cette approche, la technique du squelette-étoile est d'abord appliquée au contour de la silhouette pour obtenir des

points caractéristiques. Ensuite, ces points, ainsi que le centre de gravité de la silhouette, sont utilisés pour calculer les vecteurs de caractéristiques et les valeurs de profondeur du corps. Ces vecteurs caractéristiques sont enfin utilisés comme entrées pour un réseau de neurones à apprentissage par quantification vectorielle (Learning Vector Quantification, LVQ) pour déterminer la posture du sujet humain. Amine Elforaici et al., (2018) ont proposé deux méthodes supervisées pour la reconnaissance de posture humaine. Dans la première méthode, des réseaux de neurones profonds (CNNs) sont entraînés en utilisant l'apprentissage par transfert (transfert learning), tandis que la seconde méthode utilise la configuration des articulations du corps dans l'espace 3D pour modéliser la posture, puis effectue la reconnaissance de postures en utilisant un SVM basé sur des caractéristiques du squelette 3D. Xu et al., (2019) ont proposé une méthode de reconnaissance de posture humaine en utilisant la caméra Kinect V2. Dans cette méthode, un tracker de squelette est utilisé pour identifier les emplacements des articulations du corps humain. À partir de ces emplacements, la tâche de reconnaissance de posture est effectuée à l'aide d'un réseau de neurones basé sur l'apprentissage par rétro-propagation. Li et al., (2019) ont également utilisé un réseau de neurones en combinaison avec des données de profondeurs, des données de squelettes et des paramètres anthropométriques du corps humain. Le rapport entre la hauteur des postures des personnes et celle de leur tête est initialement calculé, puis utilisé pour distinguer entre quatre types de postures, à savoir "Standing", "Sitting", "Kneeling" et "Sitting cross-legged". Des vecteurs de caractéristiques sont ensuite extraits, transformés, puis introduits dans un réseau de neurones basé sur l'apprentissage par rétro-propagation afin de reconnaître deux autres postures, à savoir "Bending" et "Lying". Ding et al., (2020) ont tout d'abord extrait un vecteur de dimension 219 comprenant des caractéristiques d'angle et de distance entre les articulations. Ensuite, un algorithme de classification basé sur un apprentissage de règles est utilisé pour la reconnaissance de postures.

Le principal avantage des approches 3D est qu'elles sont généralement plus précises et moins sensibles aux changements du point de vue que les approches 2D. Cependant, la plupart de ces approches sont relativement complexes et nécessitent des temps de calcul importants.

4.3. Approche proposée

Dans le système que nous proposons, l'étape de reconnaissance de posture humaine est effectuée en parallèle avec l'étape de suivi. Dans l'ensemble, notre approche proposée pour la reconnaissance de posture humaine se compose de deux étapes principales, à savoir : 1) l'extraction des caractéristiques de posture, dans laquelle des caractéristiques appropriées et pertinentes caractérisant la posture humaine sont extraites; et 2) la classification de posture, dans laquelle les caractéristiques extraites sont utilisées comme entrées à un classifieur SVM multi-classe pour déterminer la classe de la personne en cours de suivi. Les détails sur chacune de ces étapes sont présentés dans les sous-sections suivantes.

4.3.1. Extraction des caractéristiques de posture

Afin de rendre notre approche insensible aux changements de couleur et de la texture des vêtements que portent les personnes en mouvement, nous avons utilisé uniquement la silhouette de la personne détectée afin d'extraire les caractéristiques décrivant sa posture. Ces caractéristiques extraites sont : les moments de Krawtchouk, l'histogramme de chaîne de codes et les caractéristiques géométriques.

4.3.1.1. Moments de Krawtchouk

Les moments de Krawtchouk (Mesbah et al., 2016; Gautam et al., 2017) sont de puissants descripteurs de forme basés sur la région, et qui ont été largement utilisés dans de nombreuses applications d'analyse d'image et de reconnaissance de formes. En raison de leur propriété d'orthogonalité, ce type de caractéristiques possède la capacité de décrire une image sans redondance ou chevauchement d'informations entre les moments. Par ailleurs, contrairement à d'autres types de moments d'image, tels que les moments de Zernike et les moments de Legendre, les moments de Krawtchouk sont directement définis dans le domaine discret, ce qui signifie que leurs calculs ne nécessitent pas de transformation dans l'espace de coordonnées de l'image. Cette propriété permet aux moments de Krawtchouk d'être plus précis et relativement moins coûteux en termes de temps de calcul.
Le calcul des moments de Krawtchouk est effectué en projetant l'image sur les polynômes de Krawtchouk. Rappelons que les polynômes de Krawtchouk d'ordre n-

ième à un point discret x peuvent être définis en termes de la fonction hypergéométrique comme suit :

$$K_n(x;p,N) = {}_2F_1\left(-n,-x;-N;\frac{1}{p}\right) \tag{4.1}$$

où, $x, n = 0,1,2,\dots,N,\ N > 0,\ p \in (0,1)$, et ${}_2F_1$ est la fonction hypergéométrique définie par :

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} \tag{4.2}$$

avec, $(a)_k$ est le symbole de Pochhammer défini comme suit :

$$(a)_k = a(a+1)(a+2)\dots(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{4.3}$$

Des exemples de polynômes de Krawtchouk jusqu'au 2-ème ordre sont : $K_0(x;p,N) = 1$; $K_1(x;p,N) = 1 - \left[\frac{1}{Np}\right]x$; $K_2(x;p,N) = 1 - \left[\frac{2}{Np} + \frac{1}{N(N-1)p^2}\right]x + \left[\frac{1}{N(N-1)p^2}\right]x^2$. Afin d'éviter les fluctuations numériques dans le calcul des moments, les polynômes de Krawtchouk sont pondérés comme suit :

$$\overline{K}_n(x;p,N) = K_n(x;p,N)\sqrt{\frac{w(x;p,N)}{\rho(n;p,N)}} \tag{4.4}$$

où, $w(x;p,N)$ et $\rho(n;p,N)$ sont, respectivement, la fonction de pondération et la norme définies dans les équations suivantes :

$$w(x;p,N) = \binom{N}{x} p^x (1-p)^{N-x} \tag{4.5}$$

$$\rho(n;p,N) = (-1)^n \left(\frac{1-p}{p}\right)^n \frac{n!}{(-N)_n} \tag{4.6}$$

Pour rendre le calcul des polynômes de Krawtchouk pondérés moins coûteux en temps, la relation de récurrence suivante peut être utilisée :

$$p(n-N)\overline{K}_{n+1}(x;p,N) = A(Np - 2np + n - x)\overline{K}_n(x;p,N) - Bn(1-p)\overline{K}_{n-1}(x;p,N) \tag{4.7}$$

avec, $A = \sqrt{\frac{(1-p)(n+1)}{p(N-n)}}$; $B = \sqrt{\frac{(1-p)^2(n+1)n}{p^2(N-n)(N-n+1)}}$, et les polynômes de Krawtchouk pondérés initiaux $\overline{K}_0(x;p,N) = \sqrt{w(x;p,N)}$, et $\overline{K}_1(x;p,N) = \left(1 - \frac{x}{pN}\right)\sqrt{w(x;p,N)}$.

Ainsi, pour une image numérique de dimension spatiale $N \times M$ avec une fonction d'intensité $f(x,y)$, les moments de Krawtchouk d'ordre $(n+m)$, en termes de polynômes de Krawtchouk pondérés, sont obtenus comme suit :

$$Q_{n,m} = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} \overline{K}_n(x;p_1,N-1)\overline{K}_m(y;p_2,M-1)f(x,y) \tag{4.8}$$

Les deux paramètres p_1 et p_2 permettent aux moments de Krawtchouk d'extraire des caractéristiques locales de n'importe quelle région d'intérêt spécifique de l'image. Leurs valeurs sont fixées à 0.5 dans nos expériences.
Enfin, notons que le nombre total, n_{Total}, de moments de Krawtchouk jusqu'à un ordre maximum n_{Max}, $0 \leq (n+m) \leq n_{Max}$, utilisé dans nos expériences pour décrire la posture humaine est donné par :

$$n_{Total} = \frac{(n_{Max}+1)(n_{Max}+2)}{2} \tag{4.9}$$

4.3.1.2. Histogramme de chaine de code

L'histogramme de chaîne de code est obtenu en calculant les fréquences d'occurrence des codes de direction présents dans la représentation de chaîne de code (CC) du contour de la silhouette humaine et ce en utilisant le descripteur illustré dans la Figure 2.11(c) du chapitre 2. Cependant, dans notre approche, avant d'extraire l'histogramme de chaîne de code, et afin de réduire l'effet du bruit et des changements d'échelle sur la représentation par une chaîne de codes, nous proposons de ré-échantillonner le contour de la silhouette humaine de manière à ce qu'il contient le même nombre de points à chaque trame. En général, dans la littérature, il existe trois méthodes distinctes pour le ré-échantillonnage d'un contour (Bourennane and Fossati, 2012): l'échantillonnage à angle égal (equal angle sampling), l'échantillonnage à points égal (equal points sampling), et l'échantillonnage à longueur d'arc égale (equal arc-length sampling). Supposons que le nombre total de points à échantillonner le long du contour est noté par N_p. La

méthode d'échantillonnage à angle égal sélectionne les points candidats espacés d'un angle égal à $\theta = 2\pi/N_p$. La méthode d'échantillonnage à points égal sélectionne les points candidats espacés d'un nombre de points égal le long du contour. L'espace entre deux points candidats consécutifs est donné par L_c/N_p, où L_c est la longueur, ou le nombre total de points, du contour. La méthode d'échantillonnage à longueur d'arc égale sélectionne les points candidats espacés par une longueur d'arc égale le long du contour. L'espace entre deux points candidats consécutifs est donné par P/N_p, où P est le périmètre du contour. Dans notre approche, la méthode d'échantillonnage à longueur d'arc égale est utilisée, car elle offre le meilleur espacement entre les points le long du contour (Bourennane and Fossati, 2012). Une illustration de cette méthode lorsqu'elle est appliquée à un exemple de contour de silhouette humaine est donnée dans la Figure 4.1. À partir de cette figure, nous pouvons observer que la méthode de ré-échantillonnage non seulement normalise la taille du contour de la silhouette, mais a également un effet de filtrage du bruit et des petits détails du contour. En faisant varier la valeur du paramètre N_p, nous pouvons ajuster la quantité de bruits et d'informations locales filtrées du contour de la silhouette et, par conséquent, la précision de la représentation du contour.

Afin d'illustrer la capacité de l'histogramme de chaîne de code dans la description de la posture humaine, quelques exemples de postures humaines avec leur histogramme correspondant sont donnés dans la Figure 4.2. Dans cette figure, la première ligne montre les silhouettes d'un humain en cinq postures différentes; la deuxième ligne montre les contours correspondants ré-échantillonnés en N_p = 32 points, tandis que la dernière ligne montre les histogrammes de chaîne de code à 8 directions obtenus en se déplaçant dans le sens des aiguilles d'une montre. À partir de cette figure, nous pouvons observer que les histogrammes de chaîne de codes des différentes postures du corps humain ont des formes globales différentes, ce qui peut être très utile pour la reconnaissance de posture humaine.

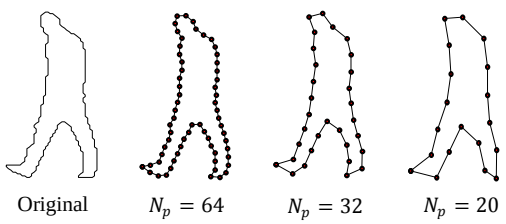


Figure 4.1: Exemple d'un contour humain ré-échantillonné en différent nombre de points, $N_p = 64, 32$ et 20, avec le contour original constitué de 267 points.

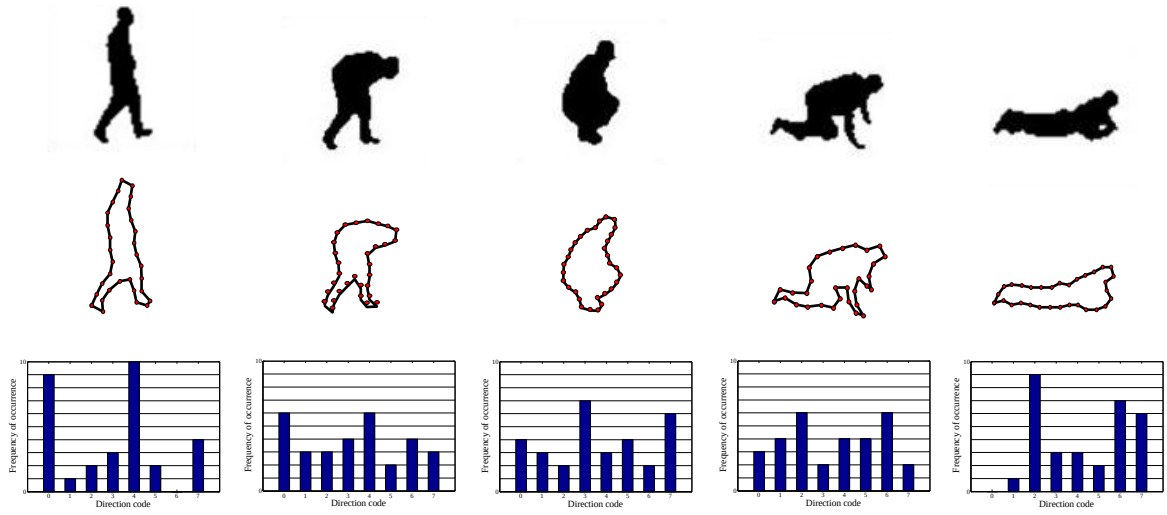


Figure 4.2: Histogrammes de chaine de code pour différentes postures du corps humain.

4.3.1.3. Caractéristiques géométriques

Les caractéristiques géométriques sont des caractéristiques simples, mais très efficaces pour distinguer entre différents types de posture humaine. Ces caractéristiques sont moins performantes lorsqu'elles sont utilisées seules, mais lorsqu'elles sont combinées, elles peuvent offrir une représentation plus compacte de la posture humaine. Dans notre approche proposée, les caractéristiques géométriques extraites pour décrire la posture du corps humain comprennent : a) le rapport de forme, b) l'inclinaison du corps humain, et c) les distances du centre de gravité de la silhouette au point de sommet le plus proche et le plus éloigné de l'enveloppe convexe. Ces caractéristiques, en plus de leur simplicité et leur faible coût de calcul, elles satisfont toutes la propriété d'invariance au changement d'échelle de la silhouette du corps humain.

4.3.1.3.1. Rapport de forme

Le rapport de forme, ou rapport d'aspect (*Aspect Ratio*, *AR*) est la caractéristique la plus simple qui peut être utilisée pour distinguer entre différents types de posture du corps humain. Cette caractéristique est souvent calculée comme le rapport entre la hauteur (H_{MBB}) et la largeur (L_{MBB}) de la boîte englobante minimale (*Minimal Bounding Box*, *MBB*) de la silhouette humaine. Cependant, dans

les situations réelles, l'utilisation de la *MBB* pour estimer la hauteur et la largeur de la silhouette humaine peut conduire à des résultats imprécis en raison du mouvement non rigide des parties du corps humain. Ainsi, dans notre approche proposée, afin d'obtenir une estimation plus précise de la hauteur et de la largeur du corps humain, nous utilisons la *MBB* de la meilleure ellipse ajustée (*Best Fitting Ellipse*, *BFE*) de l'humain détecté au lieu de la *MBB* de sa silhouette corporelle. La formule pour calculer le rapport de forme peut être ainsi exprimée comme suit :

$$AR = \frac{H_{BFE}}{L_{BFE}} \tag{4.10}$$

où, H_{BFE} et L_{BFE} sont, respectivement, la hauteur et la largeur de la *MBB* de la *BFE* de l'humain détecté.

A titre d'illustration, la Figure 4.3(a) présente une comparaison entre la *MBB* et la *MBB* de la *BFE* pour un exemple de silhouette humaine. À partir de cette figure, nous pouvons clairement observer que dans le cas de la présence d'un mouvement non rigide d'une partie du corps humain (exemple, un bras levé), la *MBB* de la *BFE* fournit une estimation plus précise de la taille humaine par rapport à la *MBB*.

4.3.1.3.2. Angle d'inclinaison

L'angle d'inclinaison du corps humain est une autre caractéristique importante qui peut être utilisée pour distinguer entre différents types de posture du corps humain. Cette caractéristique est estimée comme étant l'angle que forme l'axe principal de la *BFE* de la silhouette humaine par rapport à l'axe vertical (axe des ordonnées). La formule est donnée par l'équation suivante :

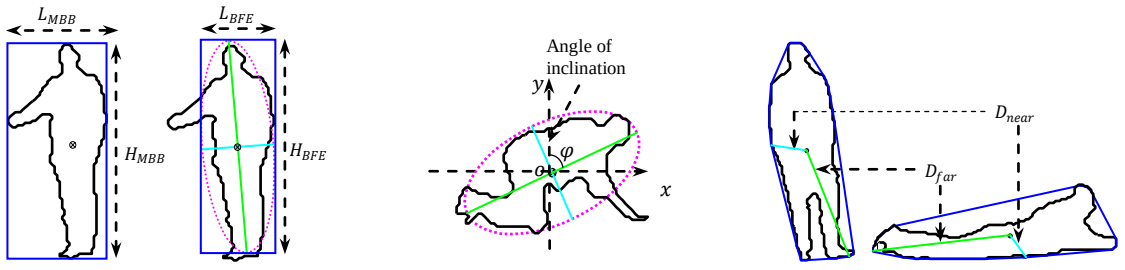
$$\varphi = 90^{\circ} - abs(\theta) \tag{4.11}$$

où, θ est l'angle défini précédemment dans l'équation 2.23 du chapitre 2. Une illustration de l'angle d'inclinaison φ pour un exemple de silhouette humaine est donnée dans la Figure 4.3(b).

4.3.1.3.3. Distances du centre de gravité de la silhouette au point de sommet le plus proche et le plus éloigné de l'enveloppe convexe

L'enveloppe convexe d'une silhouette humaine est définie comme la plus petite région convexe qui englobe tous les pixels de cette silhouette. Rappelons qu'une région est convexe si et seulement si tous les segments de ligne reliant une paire de points dans la région se trouvent également dans la région. Dans notre approche proposée, pour décrire la posture de l'humain détecté, nous extrayons deux caractéristiques différentes de l'enveloppe convexe englobant sa silhouette. Ces caractéristiques, montrées sur la Figure 4.3(c), comprennent les distances D_{near} et D_{far} du centre de gravité de la silhouette au point de sommet le plus proche et le plus éloigné de l'enveloppe convexe. Cependant, dans certaines situations réelles, deux postures humaines différentes peuvent avoir des valeurs très proches, voire identiques, pour les distances D_{near} et D_{far} , comme c'est le cas pour les deux postures ("debout" et "couché") illustrées dans la Figure 4.3(c). Pour surmonter ce problème, nous effectuons une normalisation des distances D_{near} et D_{far} par rapport à la hauteur H_{BFE} de la *MBB* de la *BFE* de la silhouette humaine détectée. La formule de calcul peut donc s'exprimer comme suit :

$$\begin{cases} \widehat{D}_{near} = \dfrac{D_{near}}{H_{BFE}} \\ \widehat{D}_{far} = \dfrac{D_{far}}{H_{BFE}} \end{cases} \tag{4.12}$$



(a) (b) (c)

Figure 4.3: Caractéristiques géométriques. (a) La *MBB* et la *MBB* de la *BFE*. (b) Angle d'inclinaison. (c) Distances du centre de gravité de la silhouette au point de sommet le plus proche et le plus éloigné de l'enveloppe convexe.

4.3.2. Classification de posture

Comme nous l'avons mentionné précédemment, chacune des trois caractéristiques décrites dans la section 4.3.1 caractérise différents aspects du contenu de la silhouette du corps humain. Ainsi, dans notre approche proposée, afin d'obtenir un résultat de reconnaissance de posture plus précis et plus fiable, nous avons tout d'abord concaténé les trois types de caractéristiques dans un seul vecteur de caractéristiques, puis nous avons utilisé le résultat comme entrée vers un algorithme SVM à base de dendogrammes afin de décider de la classe de posture de la personne suivie.

4.3.2.1. SVM basé sur les dendogrammes

Le classifieur SVM décrit dans la section 2.3.3.2.3 du chapitre 2 a été initialement conçu pour des tâches de classification binaire (c.-à-d., la discrimination entre deux classes uniquement). Cependant, comme dans de nombreux problèmes de classification, la classification de posture humaine nécessite une discrimination entre plusieurs classes de postures (Bending, Squatting, Lying, etc.). Ainsi, afin de bénéficier de la capacité discriminative du classifieur SVM binaire, plusieurs approches pour étendre ce classifieur aux problèmes de classification multi-classes ont été proposées dans la littérature. Les deux approches les plus couramment utilisées sont l'approche "Un Contre Tous" (One-Against-All, OAA) et l'approche "Un Contre Un" (One-Against-One, OAO). Dans l'approche OAA, un ensemble de K classifieurs SVM binaires (où K est le nombre de classes) est utilisé pour distinguer chaque classe de toutes les autres classes. Le résultat final de classification est obtenu en utilisant la stratégie de décision "winner-takes-all". Dans l'approche OAO, $K(K-1)/2$ classifieurs sont construits, chacun étant entraîné pour chaque paire de classes possible. La décision finale est obtenue en appliquant la stratégie du vote majoritaire (majority voting strategy). Les approches OAA et OAO ont démontré des performances supérieures sur des tâches de classification multi-classes. Cependant, ces approches nécessitent un temps de calcul relativement important, ce qui réduit la capacité de traitement en temps réel. Ainsi, dans notre approche, pour résoudre le problème de classification multi-classes, nous avons utilisé le classifieur SVM basé sur les dendogrammes (Dendogram-based Support Vectors Machines, DSVM), proposé par Benabdeslem

and Bennani, (2006), qui combine l'efficacité de calcul de la méthode de classification ascendante hiérarchique (Ascendant Hierarchical Clustering, AHC) et la haute précision de classification du SVM binaire.

Le DSVM est un classifieur SVM multi-classe basé sur les arbres de décision. Il est composé de deux étapes principales. La première consiste à calculer les K centres de gravité pour les K classes connues en utilisant les valeurs moyennes de leurs caractéristiques, puis appliquer la méthode AHC sur ces K centres. La deuxième étape consiste à associer chaque SVM binaire à un nœud, et ensuite entraîner le SVM avec les éléments des deux sous-ensembles de ce nœud. Un exemple illustrant la classification d'un échantillon de test $\boldsymbol{x}$ à l'aide du DSVM est présenté dans la Figure 4.4. La procédure AHC est illustrée par les cases en pointillées.

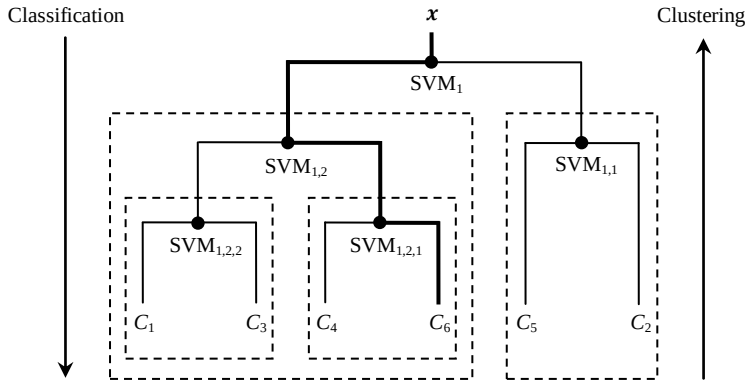


Figure 4.4: SVM basé sur les dendogrammes (DSVM) pour la classification multi-classe.

Dans cet exemple, le SVM_1 est entraîné en considérant les éléments de $\{C_2, C_5\}$ comme positifs et les éléments de $\{C_1, C_3, C_4, C_6\}$ comme négatifs ; le $SVM_{1,2}$ est entraîné en considérant les éléments de $\{C_4, C_6\}$ comme positifs et les éléments de $\{C_1, C_3\}$ comme négatifs. Ce processus est répété pour tous les SVMs binaires associés aux nœuds du dendogramme, formant ainsi $K - 1$ ($K = 6$) SVMs pour un problème à K classes. La classification d'un échantillon de test $\boldsymbol{x}$ est effectuée comme suit. Tout d'abord, $\boldsymbol{x}$ est présenté au SVM_1, puis sa sortie est $\boldsymbol{x} \in \{C_1, C_3, C_4, C_6\}$. Ensuite, $\boldsymbol{x}$ est présenté au $SVM_{1,2}$, puis sa décision de sortie est $\boldsymbol{x} \in \{C_4, C_6\}$. Enfin, $\boldsymbol{x}$ est présenté au $SVM_{1,2,1}$, et sa sortie est $\boldsymbol{x} \in C_6$. Le DSVM donne une trace de $\boldsymbol{x}$ qui est $SVM_1 \rightarrow SVM_{1,2} \rightarrow SVM_{1,2,1}$. De cet exemple, nous pouvons observer que le DSVM à 6 classes nécessite 5 SVMs binaires dans la phase d'apprentissage, mais

dans la phase de test, il ne nécessite que 3 SVMs parmi les 5 entraînés, ce qui réduit considérablement le temps de classification grâce à l'ensemble optimal des SVMs sélectionnés de manière descendante dans le dendogramme.

4.3.2.2. Filtrage de posture

Sur la base du DSVM décrit dans la section précédente, la reconnaissance de posture de la personne suivie peut être effectuée à chaque trame indépendamment des résultats obtenus dans les trames précédentes. Cependant, dans les scénarios du monde réel, reconnaître la posture humaine de cette manière est souvent moins efficace et peut entraîner un nombre important d'erreurs, en particulier lorsque le masque d'avant-plan extrait est de très mauvaise qualité. Ainsi, afin de réduire l'effet de ces erreurs sur la performance de notre approche proposée, nous avons effectué un lissage du résultat de classification de posture en utilisant l'historique récent des sorties du DSVM, et qui est obtenu à l'aide de l'information temporelle fournie par l'algorithme de suivi. Cette procédure de lissage a été menée sous l'hypothèse que, dans la plupart des cas, et pour un taux d'acquisition d'images suffisamment élevé, la posture de la personne suivie ne varie pas brusquement entre deux trames consécutives. Ainsi, la décision actuelle sur la posture humaine peut être lissée en considérant l'historique récent des sorties du classifieur DSVM. Pour effectuer cette tâche, nous utilisons le filtre à vote majoritaire pondéré (Weighted Majority Voting, WMV) exprimé par les équations suivantes :

$$O_{WMV}(t) = \underset{k}{arg\max} \left(\sum_{j=t,t-1,\dots,t-T} eq(O_{DSVM}(j), c_k) G_\sigma(t-j) \right) \tag{4.13}$$

$$eq(a,b) = \begin{cases} 1 & si\ \ a = b \\ 0 & si\ \ a \neq b \end{cases} \tag{4.14}$$

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{4.15}$$

où, O_{DSVM} et O_{WMV} sont, respectivement, les sorties du classifieur DSVM et du filtre WMV, la fonction $eq()$ est utilisée pour déterminer si les deux variables a et b ont la même valeur, $c_1, c_2, \dots, c_K$ sont les K étiquettes de classe de posture, G_σ est la fonction de pondération gaussienne avec un écart type σ, et T est la longueur (en trames) de l'historique temporel considéré.

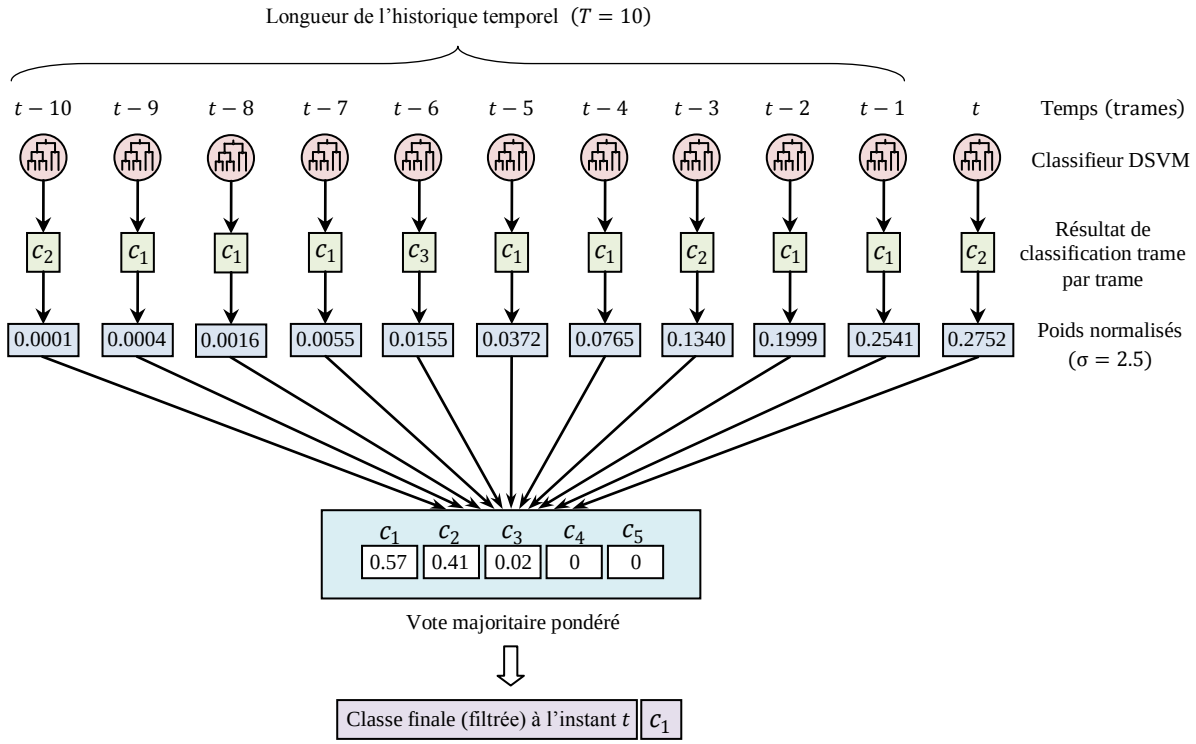


Figure 4.5: Filtrage temporel de posture humaine en utilisant le filtre WMV.

La Figure 4.5 montre le processus de filtrage temporel de la posture humaine à l'aide du filtre WMV. Le filtre associe à chaque sortie du DSVM dans l'historique temporel considéré un poids différent selon l'indice temporel (trame). Les trames les plus récentes ont plus d'importance que les trames précédentes lors de la prise de décision finale. Le résultat final (filtré) de la classification de posture à l'instant t est obtenu sur la base du vote pondéré le plus élevé.

4.4. Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche pour la reconnaissance de posture humaine à partir de séquences d'images. L'approche proposée est basée sur une combinaison de trois caractéristiques différentes, à savoir : les moments de Krawtchouk, l'histogramme de chaîne de codes et les caractéristiques géométriques. Chaque type de ces caractéristiques représente différents aspects visuels de la silhouette humaine. Les moments de Krawtchouk, par exemple, représentent la région à l'intérieur de la silhouette humaine, l'histogramme de chaîne de codes représente les points du contour qui entoure la silhouette humaine, alors que les caractéristiques géométriques décrivent

l'information globale du corps humain. Après avoir extraire ces différentes caractéristiques, elles sont concaténées pour former un seul vecteur de caractéristiques, qui est ensuite envoyé vers un SVM multi-classe à base de dendogrammes afin de décider de la classe de posture de la personne suivie. Enfin, afin de réduire l'effet des erreurs de segmentation sur la performance globale de notre système proposé, le résultat de classification de posture est filtré en utilisant la méthode du vote majoritaire pondéré et l'information temporelle fournie par l'algorithme de suivi. Dans le prochain chapitre, nous exposerons les résultats expérimentaux de l'ensemble des parties constituant notre système de vidéo surveillance intelligente proposé.

Chapitre 5

Résultats expérimentaux

5.1. Introduction

Dans ce chapitre, nous présentons les expériences menées pour évaluer les performances de notre système de vidéo surveillance intelligent proposé. Cette évaluation est menée comme suit. Dans la Section 5.2, nous décrivons les différentes bases de données utilisées pour mener les expériences. Ensuite, dans la Section 5.3, nous présentons les résultats expérimentaux obtenus par nos deux approches de détection de personnes proposées. Dans la Section 5.4, nous présentons les expériences réalisées pour évaluer l'approche de suivi proposée, qui est basée sur un filtre à particules et une combinaison adaptative de multiples caractéristiques. Dans la Section 5.5, nous présentons les expériences réalisées pour évaluer l'approche de reconnaissance de posture humaine proposée, qui est basée sur les histogrammes de chaîne de codes, les moments de Krawtchouk, et les caractéristiques géométriques. Nous évaluons les performances de ces caractéristiques individuellement, dans un premier temps, puis en combinaison afin d'étudier leurs effets séparés et combinés. Dans la Section 5.6, nous testons la performance globale de notre système proposé pour détecter un comportement humain anormal dans le contexte de vidéo surveillance nocturne en extérieur. Enfin, dans la Section 5.7, nous terminerons ce chapitre par une conclusion qui synthétisera les principaux résultats obtenus durant les différentes expériences.

5.2. Bases de données

Ci-après sont décrites en détails les différentes bases de données que nous avons utilisées pour réaliser nos expériences.

5.2.1. Base de données de postures humaines

Vu qu'à notre connaissance, il n'existe pas de bases de données publiques dédiées à la détection de personnes et à la reconnaissance de leurs postures dans des séquences d'images IR, nous avons décidé de construire notre propre base de données de postures humaines que nous utiliserons pour évaluer notre système proposé. Dans le but de réaliser cette tâche, nous avons utilisé le système d'acquisition illustré dans la Figure 5.1 afin d'acquérir un ensemble de séquences d'images IR. Ce système comporte trois éléments principaux, à savoir : une caméra proche IR, un Convertisseur Analogique/Numérique (CAN), et un ordinateur. Chacun de ces éléments effectue une tâche bien précise, qui est décrite ci-dessous.
La caméra proche IR, de type CCD, est utilisée pour l'acquisition des séquences d'images dans un environnement extérieur. Quelques caractéristiques principales de cette caméra sont présentées dans le Tableau 5.1. Grâce à son illuminateur infrarouge intégré comprenant 42 LEDs proche IR (840~850 nm), cette caméra est capable de capturer des images en niveau de gris de la scène observée dans des conditions d'obscurité totale, et ce, jusqu'à une distance maximale de 40 mètres.
Comme la caméra proche IR utilisée est une caméra analogique, un CAN est nécessaire afin de convertir le signal vidéo analogique délivré par la caméra en un signal numérique exploitable par la machine. Dans notre cas, un TV Box & Grabber USB 2.0 est utilisé afin d'effectuer la conversion analogique/numérique. Ce convertisseur est tout d'abord relié à la caméra via un câble coaxial avec un connecteur de type BNC, puis il est connecté à un ordinateur via un câble USB 2.0 pour une utilisation ultérieure des séquences d'images acquises.
Les caractéristiques techniques de l'ordinateur utilisé pour l'acquisition et le traitement des séquences d'images IR sont données dans le Tableau 5.2.

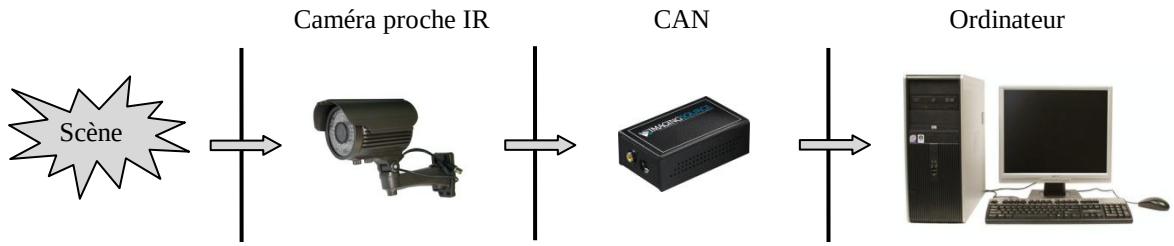


Figure 5.1 : Les éléments constituant le système d'acquisition.

Tableau 5.1: Quelques caractéristiques de la caméra proche IR utilisée pour l'acquisition des séquences d'images.

Spécification	**Valeur**
Nom du modèle	Sunell SN-IPR54/12DN
Résolution maximale	1080p (1920×1080)
Type de capteur	CCD
Illuminateur IR	42 LEDs (840~850 nm)
Portée des IRs	Jusqu'à 40 m
Fréquence de capture	25 fps

Tableau 5.2: Quelques caractéristiques de l'ordinateur utilisé dans les expérimentations.

Composante	**Valeur**
Processeur	**Intel® Core™ i5**
Fréquence de la CPU	**2.80 GHz**
Mémoire RAM	**4 Go**
Carte graphique	NVIDIA GeForce 9400 GT
Système d'exploitation	Windows 7 Ultimate

En suivant la procédure d'acquisition décrite dans la Figure 5.1, nous avons capturé un ensemble de séquences d'images proche IR dans un environnement extérieur et dans des conditions nocturnes réalistes. La caméra proche IR a été placée à l'extérieur d'une maison privée, à environ trois mètres du sol, et inclinée vers le bas à environ 20-25 degrés de l'horizontale. Les séquences d'images ont été acquises avec une résolution de 352×288 pixels, et à une fréquence de 25 trames par seconde. Pendant l'acquisition, nous avons demandé à deux sujets humains

adultes de simuler les postures de base suivantes : "Standing" (Debout), "Bending" (Penché), "Squatting" (Accroupi), "Creeping" (Rampant), et "Lying" (Allongé). Afin de tester la robustesse de notre système proposé face aux changements d'angle de vue, nous avons également demandé aux sujets humains de simuler les cinq postures dans huit directions de mouvement différentes par rapport à la position de la caméra. Ces directions, illustrées dans la Figure 5.2, comprennent 0°, 45°, 90°, 135°, 180°, 225°, 270° et 315°.

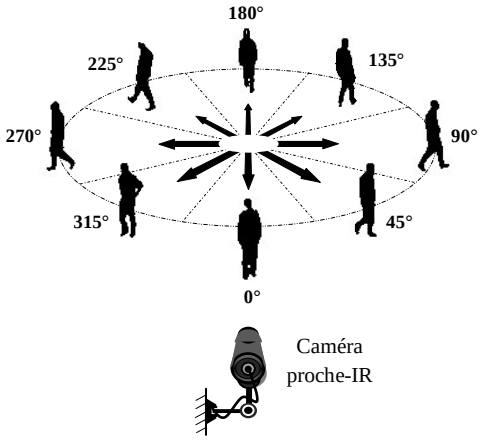


Figure 5.2: Directions de mouvement utilisées pour la capture des postures humaines (direction 0° correspond au déplacement de la personne vers la caméra).

Après avoir extrait les silhouettes humaines des séquences d'images IR en utilisant la méthode décrite dans les Sections 2.3.1.1 du Chapitre 2, nous avons sélectionné aléatoirement un nombre total de 2000 images de posture comme base de données. Pour chaque direction de mouvement, nous avons pris 50 images échantillons, ce qui correspond à un nombre total de 400 images par groupe de posture.
Quelques exemples d'images de chaque groupe de posture humaine contenues dans notre base de données créée sont présentés dans la Figure 5.3.

Groupe de posture	Direction de mouvement							
	0°	45°	90°	135°	180°	225°	270°	315°
Standing								
Bending								
Squatting								
Creeping								
Crawling								

Figure 5.3: Exemples d'images de postures humaines de notre base de données.

5.2.2. Base de données d'animaux de Bai et al.

La base de données d'animaux de Bai et al., (2009) provient d'un projet de classification de formes de l'université de science et de technologie de Huazhong en Chine. Cette base de données se compose de 20 catégories d'animaux, à savoir : Oiseau, Papillon, Chat, Vache, Cerf, Dauphin, Canard, Éléphant, Crocodile, Poisson, Oiseau volant, Poule, Cheval, Léopard, Singe, Souris, Araignée, Tortue, Lapin et Chien. Toutes ces silhouettes sont obtenues à partir d'images réelles présentant des situations difficiles telles que des auto-occultations et des variations de pose, de taille et d'angle de vue. Dans ce travail, et comme nous nous intéressons aux objets en mouvement dans le contexte de vidéo surveillance, nous avons utilisé uniquement 12 catégories de la base de données d'animaux de Bai et al., (2009) pour évaluer la performance de notre système proposé. Ces catégories sont : Chat, Vache, Cerf, Canard, Éléphant, Crocodile, Poule, Cheval, Léopard, Singe, Lapin, et Chien. Quelques exemples d'images de silhouettes provenant de ces catégories d'animaux sont donnés dans la Figure 5.4.

Groupe d'animaux	Exemples d'images
Chat	
Chien	
Vache	
Cheval	
Poule	
Lapin	
Cerf	
Canard	
Crocodile	
Léopard	
Singe	
Éléphant	

Figure 5.4: Exemples d'images de silhouettes contenues dans la base de données d'animaux de Bai et al., (2009).

5.2.3. Base de données MCL

En plus des silhouettes de personnes et d'animaux contenues dans les deux bases de données décrites précédemment, nous avons ajouté un ensemble de silhouettes de véhicules sélectionné à partir de la base de données MCL (Lee et al., 2014). Les véhicules sont capturés dans le champ de vision avec des directions et des vitesses différentes. Quelques exemples d'images de véhicules provenant de cette base de données sont donnés dans la Figure 5.5.

Figure 5.5: Exemples d'images de silhouettes de véhicules contenues dans la base de données MCL (Lee et al., 2014).

5.2.4. Base de données AIC

La base de données AIC (Conaire et al., 2006) est une collection de deux séquences d'images destinées à la détection et au suivi d'objets en mouvement dans le contexte de vidéo surveillance nocturne. Ces séquences ont été capturées depuis un balcon du campus de l'université de Dublin City, en Irlande. L'une des séquences a été capturée dans le spectre visible à l'aide d'une caméra couleur (Panasonic WV-CP470), et l'autre a été capturée dans le spectre infrarouge à l'aide d'une caméra IR thermique (Raytheon ControlIR 2000B) sensible aux longueurs d'onde de 7µm-14µm (IR long, LWIR). Les deux séquences d'images ont été enregistrées durant la nuit avec une fréquence de capture de 25 trames par seconde et une résolution d'image de 320×240 pixels. Quelques exemples de trames des séquences d'images de cette base de données sont illustrés dans la Figure 5.6. Dans nos expériences, seule la séquence d'images capturée dans le spectre infrarouge a été utilisée.

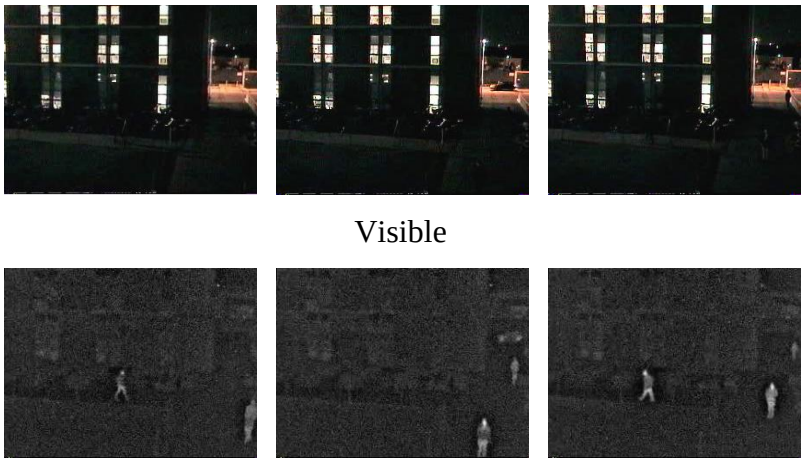


Figure 5.6: Exemples de trames de la base de données AIC (Conaire et al., 2006).

5.2.5. Base de données OTCBVS

La base de données OTCBVS (http://vcipl-okstate.org/pbvs/bench), de l'université d'État de l'Ohio (États-Unis), est une base de données de référence accessible au public pour tester et évaluer de nouvelles techniques et des algorithmes de pointe en vision par ordinateur. Elle contient des séquences d'images enregistrées dans et au-delà du spectre visible (VIS, NIR et LWIR). La principale caractéristique de cette base de données est qu'elle couvre une variété de conditions et de défis à relever, tels que des arrière-plans encombrés, des occultations, des variations d'apparence, et des changements d'échelle. L'ensemble des séquences de cette base de données est regroupé en plusieurs sous-ensembles nommés : OSU Thermal Pedestrian Database, IRIS Thermal/Visible Face Database, OSU Color-Thermal Database, Terravic Facial IR Database, Terravic Motion IR Database, Terravic Weapon IR Database et CBSR NIR Face Dataset. Dans nos expériences, des séquences des sous-ensembles OSU Color-Thermal Database (Davis and Sharma, 2007) et Terravic Motion IR Database (Miezianko and Pokrajac, 2008) sont utilisées. Quelques exemples de trames des séquences de ces sous-ensembles de données sont illustrés dans la Figure 5.7. Les séquences ont été capturées en utilisant deux caméras IR thermiques différentes : Raytheon PalmIR 250D et Raytheon L-3 Thermal-Eye 2000AS. Toutes les séquences d'images ont été prises avec une fréquence de capture de 30 trames par seconde et une résolution d'image de 320×240 pixels codés en 8 bits (niveaux de gris).

Figure 5.7: Exemples de trames des séquences de la base de données OTCBVS.

5.3. Evaluation des approches de détection de personnes

5.3.1. Première approche

Avant d'évaluer les performances de notre première approche proposée sur des séquences réelles, et afin de déterminer les "meilleures" valeurs pour le

paramètre M utilisé pour le filtrage du contour (Figure 2.3) et le paramètre θ_{legs} utilisé pour la localisation des jambes (Figure 3.6), nous avons effectué des tests expérimentaux préliminaires afin d'étudier l'évolution de la performance du descripteur squelette-étoile en fonction de chaque paramètre. Ces tests ont été effectués sur un ensemble de données contenant des échantillons positifs et négatifs. Les échantillons positifs (classe humain) comprennent des silhouettes d'humains (uniquement en postures "Standing" et "Bending") de notre base de données. Quant aux échantillons négatifs (classe non humain), ils sont constitués de l'union des silhouettes d'animaux provenant de la base de données de Bai et al., (2009) et des silhouettes de véhicules provenant de la base de données MCL (Lee et al., 2014). Les résultats obtenus en faisant varier la valeur du paramètre M à partir de l'ensemble $\{1, 2, 3, \ldots, 20\}$ et l'angle θ_{legs} à partir de l'ensemble $\{40°, 50°, 60°, \ldots, 90°\}$ sont présentés dans la Figure 5.8.

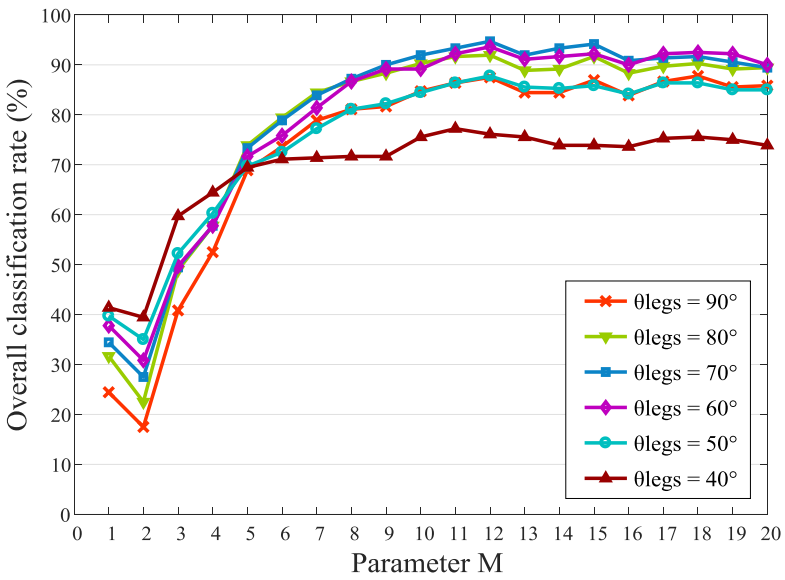


Figure 5.8: Performances du descripteur squelette-étoile pour différentes valeurs des paramètres M et θ_{legs}.

D'après la Figure 5.8, nous pouvons observer que l'augmentation de la valeur du paramètre M jusqu'à 12 améliore significativement les performances du descripteur squelette-étoile, mais l'augmentation de sa valeur au-delà de 12 diminue légèrement ses performances. D'après cette figure, nous pouvons également observer que l'augmentation de la valeur de l'angle θ_{legs} de 40° à 70° améliore

considérablement les performances du descripteur squelette-étoile, mais l'augmentation de θ_{legs} au-delà de 70° diminue ses performances. Ainsi, les valeurs $(12, 70°)$ sont utilisées comme valeurs "optimales" pour les paramètres (M, θ_{legs}) dans notre première approche de détection de personnes proposée. La matrice de confusion correspondant à cette configuration est donnée dans le Tableau 5.3. À partir de cette matrice de confusion, nous pouvons observer que le descripteur squelette-étoile permet une très bonne discrimination entre les classes humain et non humain, produisant une précision globale de classification de 94.78%.

Tableau 5.3: Matrice de confusion (en %) correspondant à la configuration M = 12 et θ_{legs} = 70°

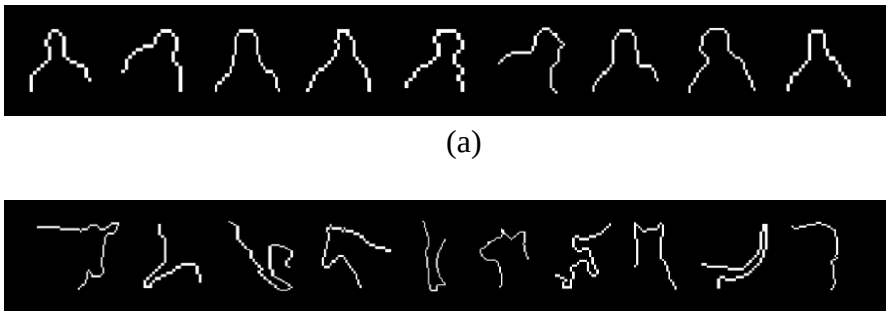
	Classe prédite	
Classe réelle	Humain	Non humain
Humain	**95.00**	05.00
Non humain	05.49	**94.51**
Précision globale : **94.78%**		

Les résultats de l'application de notre première approche sur des séquences d'images infrarouges réelles seront présentés dans la Section 5.4.

5.3.2. Deuxième approche

Afin de construire les ensembles d'apprentissage et de tests pour l'évaluation des performances de notre deuxième approche proposée, nous avons sélectionné aléatoirement, à partir de notre base de données, 10 images par groupe de postures et par direction de mouvement, ce qui correspond à un total de 10×(5 postures)×(8 directions de mouvement) = 400 images. Ces images sont utilisées pour entraîner le classifieur SVM, tandis que les autres sont laissées pour la phase de test. À partir de l'ensemble d'apprentissage construit, nous avons tout d'abord suivi la procédure décrite dans la Section 2.3.3.2.1 (Chapitre 2) pour extraire les contours candidats tête-épaules. Ensuite, nous avons étiqueté manuellement les vrais contours tête-épaule humains comme échantillons positifs, tandis que les autres contours extraits (c'est-à-dire, ceux qui ne correspondent pas à des vrais contours tête-épaule humains) sont étiquetés comme négatifs. Certains échantillons de contours étiquetés comme positifs sont montrés dans la Figure 5.9(a). Cependant, en plus

des contours qui ne correspondent pas à des parties tête-épaules humaines extraites de notre base de données, nous avons également recueilli un deuxième ensemble de contours extraits à partir de la base de données de Bai et al., (2009) et la base MCL (Lee et al., 2014). Des exemples de ces contours, étiquetés comme négatifs, sont montrés dans la Figure 5.9(b).



(a)

(b)

Figure 5.9: Exemples d'échantillons utilisés pour l'apprentissage du SVM. (a) Échantillons positifs (contours tête-épaule humains). (b) Échantillons négatifs.

Tout comme la première approche que nous avons proposée, notre deuxième approche dépend aussi des valeurs des paramètres M et θ_{legs} . Ainsi, afin de déterminer la meilleure configuration, nous avons mené une série d'expériences dans lesquelles nous avons fait varier la valeur de M de 2 à 20 (avec un pas de 2), et celle de θ_{legs} de 10° à 90° (avec un pas de 10°). Les résultats obtenus sont illustrés sur la Figure 5.10. Notons que, ces résultats sont obtenus avec un classifieur SVM linéaire et un paramètre de régularisation C = 1. D'après ces résultats, nous pouvons observer que l'augmentation de la valeur θ_{legs} jusqu'à 50° améliore significativement les performances de détection, mais au-delà de cette valeur, les performances de détection diminuent. La meilleure valeur pour le paramètre M pour la plupart des valeurs de θ_{legs} testées est égale à 12. Ainsi, dans notre deuxième approche proposée, les valeurs $(12, 50°)$ sont utilisées comme valeurs "optimales" pour les paramètres (M, θ_{legs}), fournissant un taux de précision global d'environ 88.64%.

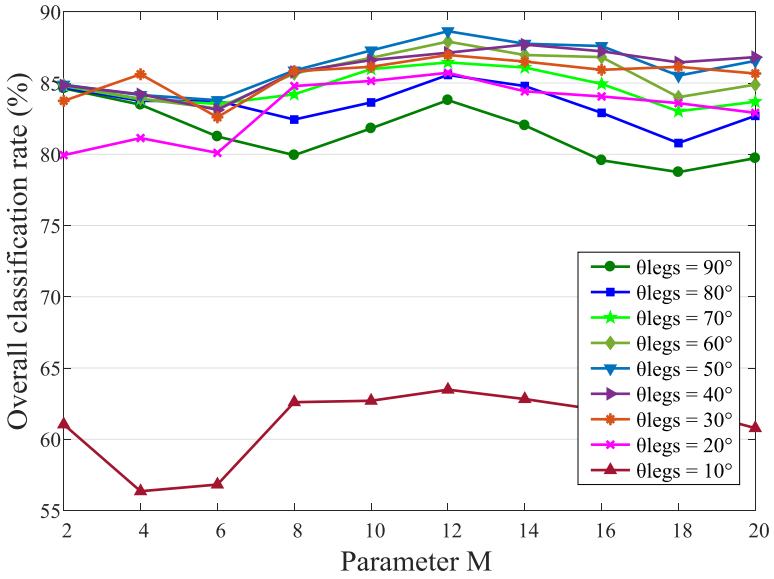


Figure 5.10: Performances de notre deuxième approche proposée pour différentes valeurs des paramètres M et θ_{legs}.

Cependant, comme les performances de notre deuxième approche proposée dépendent également du type de la fonction noyau utilisée dans le classifieur SVM, nous avons mené d'autres expériences dans lesquelles nous avons testé différentes fonctions noyaux, à savoir : linéaire, polynomiale, fonction de base radiale (Radial Basis Fonction, RBF), et le perceptron multicouche (Multi-Layer Perceptron, MLP). Dans ces expériences, et afin de trouver les meilleures valeurs pour les paramètres des noyaux, nous avons effectué la procédure de validation croisée (Bishop, 2006), au cours de laquelle plusieurs valeurs pour ces paramètres sont testées. Les valeurs qui donnent les meilleures performances de classification sont ensuite utilisées comme valeurs "optimales" dans le classifieur SVM final. Les résultats de comparaison entre les différentes fonctions noyaux sont présentés dans le Tableau 5.4. À partir de ce tableau, nous pouvons constater que, parmi toutes les fonctions, le noyau linéaire et RBF produisent les meilleures performances en termes de précision, avec 88.64% et 87.39%, respectivement. Cependant, les résultats montrent que le noyau linéaire réalise le meilleur compromis entre la sensibilité et la spécificité avec, respectivement, 88.75% et 88.12%, contre 90% et 74.37% pour le noyau RBF. Ainsi, dans le reste de nos expériences, nous adoptons la fonction linéaire dans le classifieur SVM.

Tableau 5.4: Comparaison entre différentes fonctions noyaux (paramètre de régularisation $C = 1$).

Fonction noyau	Paramètres	Sensibilité (%)	Spécificité (%)	Précision (%)
Linéaire	-	88.75	88.12	88.64
Polynomial	$d = 9$	87.81	61.56	83.43
RBF	$\sigma_{rbf} = 100$	90	74.37	87.39
MLP (Sigmoid)	$\gamma = 3,\ \delta = -2$	89	58.75	83.95

$$\text{Sensibilité} = \frac{VP}{VP+FN} \qquad \text{Specificité} = \frac{VN}{VN+FP} \qquad \text{Précision} = \frac{VP+VN}{VP+FP+VN+FN}$$

VP : Vraies Positives, VN : Vraies Négatives, FP : Fausses Positives, FN : Fausses Négatives

Afin de montrer la robustesse de notre deuxième approche proposée face aux changements de posture et de direction de mouvement de l'humain observé, nous présentons également dans la Figure 5.11 les taux de précision détaillés obtenus pour toutes les postures et directions de mouvement contenues dans notre base de données créée. De cette figure, nous pouvons observer que, parmi les cinq postures considérées, la posture "Standing" est la plus adéquate pour détecter un humain avec notre approche, alors que les postures "Creeping" et "Crawling" sont les plus difficiles. Quant aux changements de direction du mouvement, nous pouvons observer que la direction 0° (mouvement vers la caméra) est celle qui cause le plus d'erreurs pour notre deuxième approche proposée. Cela peut s'expliquer par la difficulté à détecter la partie tête-épaule dans cette direction, en particulier lorsque l'être humain observé est en mouvement dans la posture "Bending", "Creeping" ou "Crawling" (Figure 5.3).

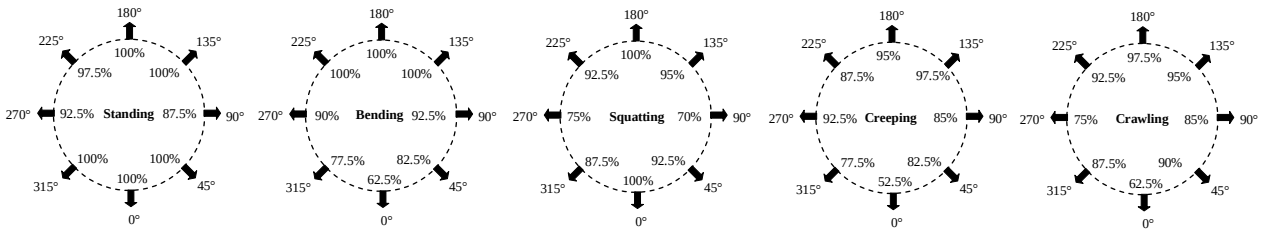


Figure 5.11: Résultats de notre deuxième approche pour différentes postures et directions de mouvement.

Enfin, pour démontrer l'avantage d'incorporer la phase de détection des jambes dans notre deuxième approche proposée, nous montrons dans la Figure 5.12 une comparaison des résultats obtenus avec et sans la prise en compte de la phase de détection des jambes. Sur cette figure, nous pouvons voir que l'incorporation de la détection des jambes améliore significativement les taux de reconnaissance de la plupart des objets non humains. Cependant, une légère diminution des taux de reconnaissance d'un humain dans les postures "Squatting" et "Creeping" a été observée. La raison de cette diminution peut être expliquée par le fait que, lorsqu'un humain se déplace dans l'une de ces posture, en particulier dans les directions 0° et 180°, sa silhouette correspondante semble avoir plus de deux jambes, et ce, en prenant en compte les bras, qui sont souvent très proches ou touchent le sol et apparaissent comme des jambes humaines supplémentaires.

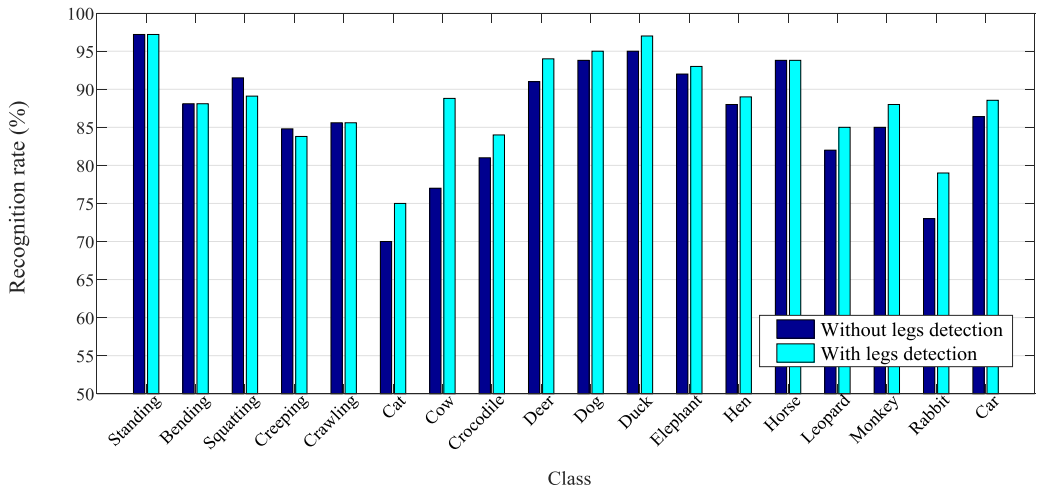


Figure 5.12: Comparaison des performances de notre deuxième approche sans et avec l'étape de détection des jambes.

5.4. Evaluation de l'approche de suivi

Dans cette section, nous présentons les différentes expérimentations que nous avons réalisées dans le but d'évaluer les performances de notre approche de suivi proposée dans le Chapitre 3. Durant ces expérimentations, et comme méthode de détection d'humains, nous avons adopté notre première approche proposée dans la Section 2.3.2. Pour mener à bien l'évaluation, nous avons utilisé un ensemble composé de six séquences d'images IR différentes tirées des deux bases de données AIC et OTCBVS présentées précédemment. Chacune de ces séquences contient des

situations et des défis différents à relever, tels que des encombrements d'arrière-plan, un fort bruit, des changements d'apparence, des occultations, des changements d'échelle, et l'apparition et la disparition de plusieurs objets en mouvement dans la scène.

Afin de valider notre approche de suivi proposée, nous avons mené une étude comparative avec quatre méthodes issues de la littérature. Ces méthodes incluent : ASLA (Adaptive Structural Local-sparse Appearance) (Jia et al., 2012), L1-APG (L1-Tracker using Accelerated Proximal Gradient) (Bao et al., 2012), Circulant Structure Kernel (CSK) (Henriques et al., 2012), Tracking using Gaussian Processes Regression (TGPR) (Gao et al., 2014). Ces méthodes sont génériques et elles peuvent être appliquées pour suivre n'importe quel objet cible dans une séquence d'images, y compris des objets non rigides et articulés tels que des humains. Les résultats expérimentaux de ces méthodes sont obtenus en exécutant le code source fourni par les auteurs sur leur propre page web. En plus de ces quatre méthodes de la littérature, nous avons également mené une étude comparative avec deux autres méthodes basées sur l'algorithme conventionnel de filtrage à particules (Nummiaro et al., 2003). L'une de ces méthodes utilise uniquement l'histogramme d'intensité, tandis que l'autre utilise uniquement l'histogramme de texture. Pour simplifier les notations, nous noterons ces méthodes par PF+Int et PF+Tex, respectivement.

Afin de comparer quantitativement les performances des différentes méthodes, nous avons utilisé trois mesures d'évaluation distinctes, à savoir : l'erreur de localisation du centre (Center Location Error, CLE), le taux de réussite (Success Rate, SR) et la vitesse de suivi en trames par second (Frame Per Second, FPS). Le CLE est calculé comme étant la distance Euclidienne entre le centre estimé de l'humain suivi et le centre de la vérité terrain (marquée manuellement). Cette mesure montre à quel point l'emplacement estimé de l'humain suivi est proche de la vérité terrain dans chaque trame. Le SR est défini comme étant le pourcentage de trames où le ratio de chevauchement ρ entre la vérité terrain et les boîtes englobantes estimées de l'humain suivi est supérieur à un seuil variant de 0 à 1. Le ratio de chevauchement ρ est défini comme suit :

$$\rho = \frac{area\,(BB_E \cap BB_{GT})}{area\,(BB_E \cup BB_{GT})} \tag{5.1}$$

où BB_E et BB_{GT} sont les boîtes englobantes estimées et de la vérité terrain, et les opérateurs ∩ et ∪ représentent l'intersection et l'union, respectivement.
Toutes les expérimentations menées dans cette partie ont été implémentées sous l'environnement MATLAB, sur un PC standard avec un processeur Intel Pentium i5 de 2.8 GHz et 4 Go de mémoire. Une liste des valeurs des paramètres utilisés dans ces expérimentations est donnée dans le Tableau 5.5.

Tableau 5.5: Liste des valeurs des paramètres utilisés dans nos expérimentations.

Paramètre	**Description**	**Valeur**
K	Nombre de distributions dans le GMM	3
τ	Seuil pour le modèle d'arrière-plan	0.2
M	Nombre de premiers coefficients de Fourier d'ordre inférieur/d'ordre supérieur	12
θ_{legs}	Intervalle d'angle pour la détection des jambes	70°
$[AR_{Min},\ AR_{Max}]$	Rapports de forme minimum et maximum pour un humain dans la scène	[1, 4]
T_{Sim}	Seuil de similarité pour la détection d'un humain	1.5
n_f	Nombre de trames consécutives pour initialiser le suivi	15
N_s	Nombre de particules dans le filtre à particules	70
m_1	Nombre de *bins* dans l'histogramme d'intensité	32
m_2	Nombre de *bins* dans l'histogramme de texture RLBP	24
σ	Paramètre modélisant le bruit dans les observations	0.15
k_0	Nombre de trames pour la mise à jour du modèle de vélocité de mouvement	90
α_{Max}	Valeur maximale des taux de mise à jour $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$	0.1
d_{Th}	Seuil de distance pour la détection d'une occultation inter-humain	10

5.4.1. Séquence 1: Encombrements d'arrière-plan

La principale difficulté dans cette séquence est l'existence d'un fort encombrement (clutter) dans l'arrière-plan de la scène, et qui est causé par le mouvement des branches et des feuilles des arbres sous l'influence du vent. La Figure 5.13(a) présente quelques trames extraites de cette séquence. La Figure 5.13(b) montre le résultat de l'étape d'extraction d'avant-plan (objets en mouvement), et la Figure 5.13(c) montre le résultat des étapes de détection et de suivi d'humain. Comme nous pouvons l'observer sur les figures, toutes les régions d'avant-plan causées par l'encombrement dans l'arrière-plan sont rejetées avec

succès par notre méthode, car elles ne satisfont pas les conditions appliquées sur la mesure de similarité globale définie dans l'équation 2.18 (Chapitre 2).

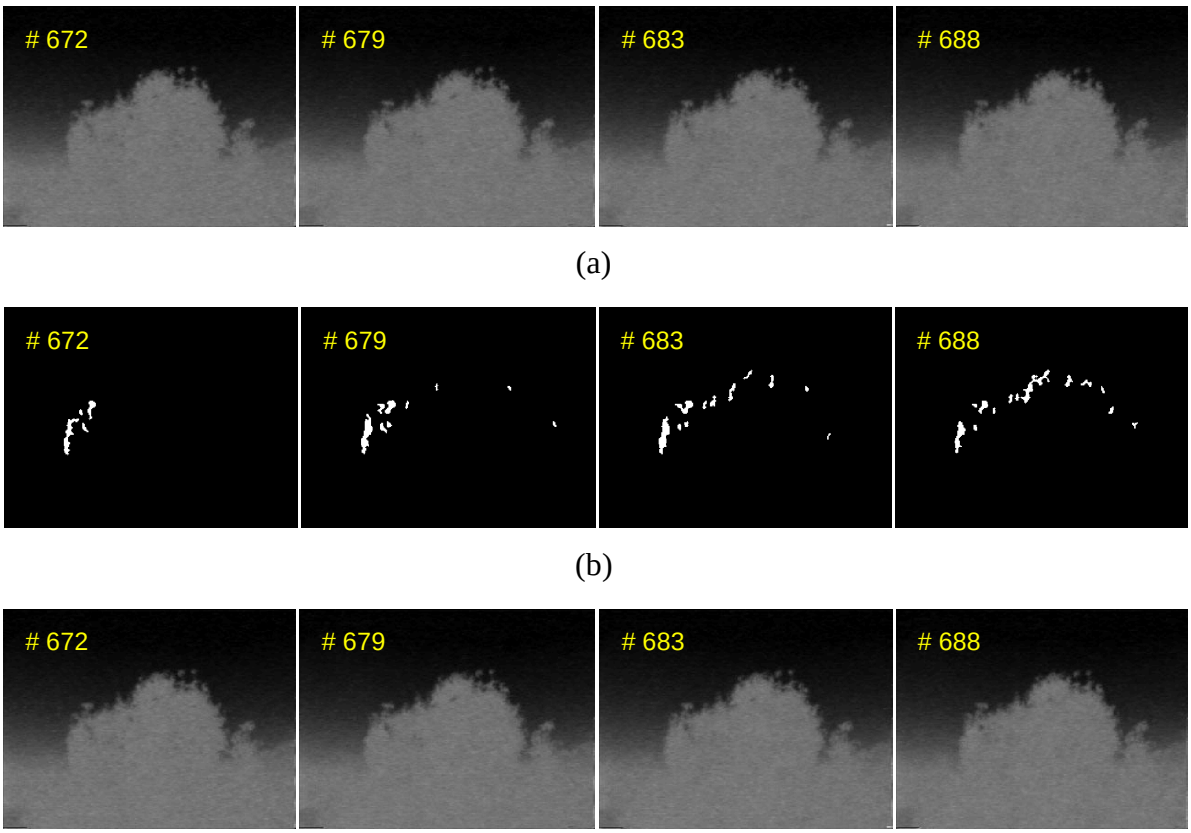
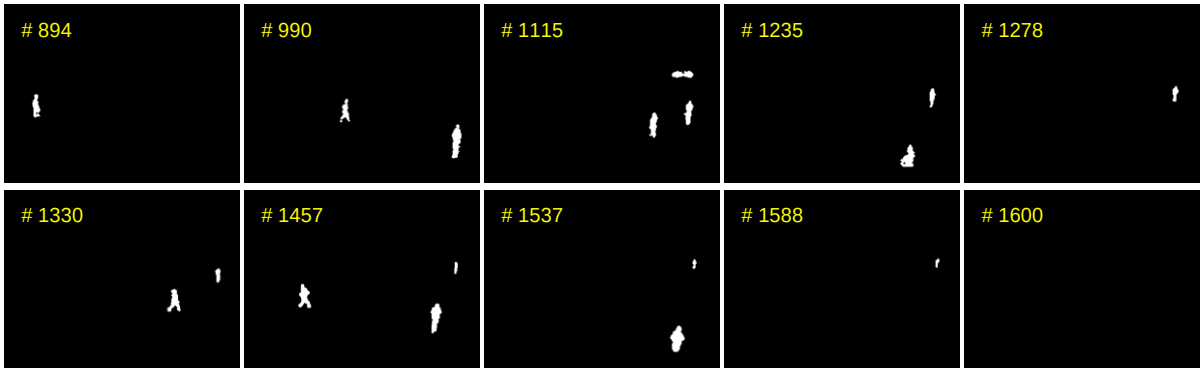


Figure 5.13: Résultats expérimentaux sur la Séquence 1 (encombrement d'arrière-plan). (a) Quelques images échantillons de la Séquence 1. (b) Résultats de l'étape d'extraction d'avant-plan (objets en mouvement). (c) Résultats des étapes de détection et de suivi d'humain.

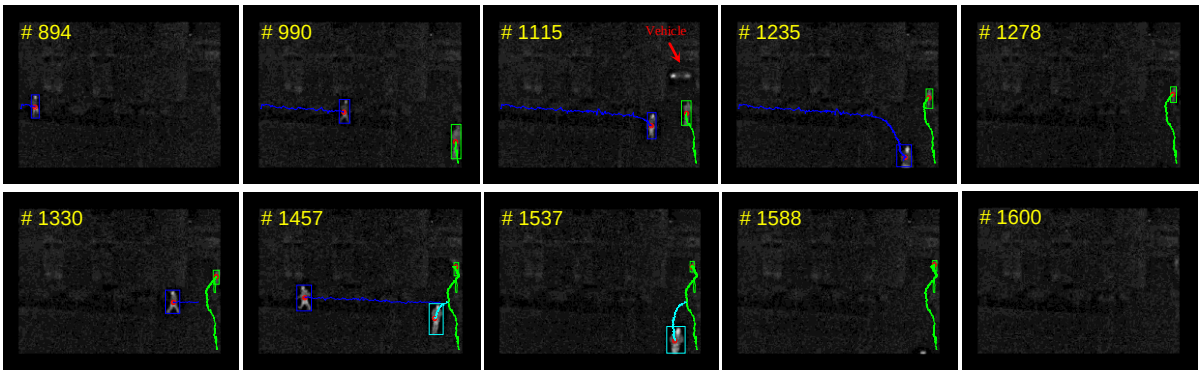
5.4.2. Séquence 2 : Apparition et disparition de plusieurs objets en mouvement

Cette séquence provient de la base de données AIC (Conaire et al., 2006), et ses principaux défis sont le faible contraste, le bruit important de la caméra, et l'apparition et la disparition de plusieurs objets en mouvement dans la scène. Quelques trames représentatives de cette séquence montrant les résultats de l'étape d'extraction d'objets en mouvement et les résultats des méthodes de détection et de

suivi d'humains proposées sont illustrées dans les Figures 5.14(a) et 5.14(b), respectivement. Les résultats des différents filtres à particules (trackers) sont représentés par des trajectoires de couleurs différentes. Le cercle rouge à l'intérieur des boîtes englobantes indique la position actuelle estimée de l'humain suivi.



(a)



(b)

Figure 5.14: Résultats expérimentaux sur la Séquence 2 (apparition et disparition de plusieurs objets en mouvement). (a) Résultats de l'étape d'extraction d'objets en mouvement. (b) Résultats des méthodes de détection et de suivi d'humain proposées.

D'après les résultats de la Figure 5.14, nous pouvons observer qu'aux alentours de la trame #894 de la séquence, un premier humain en mouvement apparaît du côté gauche du champ de vision de la caméra, et juste après quelques trames, il est efficacement détecté par notre méthode. Ainsi, un filtre à particules, marqué par une couleur bleue, est automatiquement créé pour suivre cet humain à travers la scène surveillée. Aux alentours de la trame #990, un deuxième humain apparaît

dans la scène depuis le coin inférieur droit, et une fois encore, un nouveau filtre à particules, marqué par une couleur verte, est créé pour le suivre. Aux alentours de la trame #1115, nous observons que le véhicule, indiqué par la flèche rouge, et qui apparaît du côté droit de la scène est détecté par l'étape d'extraction d'objets en mouvement, mais d'après les résultats de notre méthode, nous observons qu'il est bien rejeté par l'algorithme de suivi. Aux alentours de la trame #1235 à la trame #1278, le premier humain détecté disparaît de la scène par le côté inférieur droit, et par conséquent, son suivi est automatiquement interrompu, et le filtre à particules correspondant est réinitialisé. Aux alentours de la trame #1330, un troisième humain apparaît du côté droit de la scène, et quelques trames après, le tracker précédemment interrompu commence à le suivre le long de la scène. Aux alentours de la trame #1457, un quatrième humain apparaît dans la scène depuis le côté droit, et un nouveau filtre à particules, marqué par une couleur cyan, est à nouveau initialisé par notre méthode pour le suivre le long de la scène. Aux alentours de la trame #1537, le troisième humain détecté a complètement quitté la scène, et par conséquent, la piste correspondante est automatiquement supprimée de la liste des pistes en cours. Aux alentours de la trame #1588, le quatrième humain détecté a complètement quitté la scène par le côté inférieur droit, et par conséquent, sa piste est supprimée et le tracker correspondant est réinitialisé. Enfin, aux alentours de la trame #1600, le deuxième humain détecté, marqué par une couleur verte, est prédit hors du plan de l'image, et par conséquent, sa piste est supprimée et le filtre à particules correspondant est réinitialisé pour suivre un autre humain entrant dans la scène dans les trames ultérieures.

Afin d'illustrer comment notre méthode proposée a initialisé le suivi de l'humain dans la trame #894 et a rejeté le véhicule dans la trame #1115, nous montrons dans la Figure 5.15 la variation de la mesure de similarité globale S_{Global} , définie dans l'équation 2.18 (Chapitre 2), pour ces deux objets en mouvement sur les n_f = 15 trames consécutives suivant leur première apparition dans le champ de vision de la caméra. Comme nous pouvons l'observer sur la figure, les valeurs de la mesure de similarité globale pour l'être humain sont toutes supérieures au seuil de similarité prédéfini T_{Sim} (fixé à 1.5) pendant les n_f trames consécutives suivant sa première apparition dans la scène. En conséquence, un tracker est initialisé par notre méthode pour le suivre le long de la scène. Pour le véhicule, nous pouvons observer que sa mesure de similarité globale est supérieure au seuil T_{Sim} pendant

seulement les trois premières trames suivant sa première apparition dans la scène. Après ces trames, la valeur de la mesure de similarité globale tombe à zéro. En effet, pendant les trois premières trames, seule la partie avant du véhicule est visible par la caméra mais, après ces trames, la quasi-totalité du corps du véhicule apparaît dans le champ de vision de la caméra. Par conséquent, comme les valeurs de la mesure de similarité globale pour le véhicule ne satisfont pas à la condition selon laquelle elles doivent être supérieures au seuil T_{Sim} pendant toutes les n_f trames consécutives, le véhicule est donc rejeté par notre méthode et aucun filtre à particules n'est créé pour le suivre.

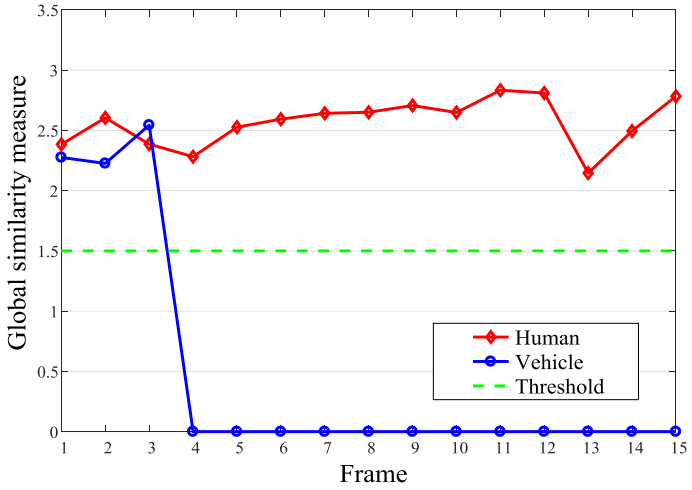


Figure 5.15: Variation de la mesure de similarité globale S_{Global} pour l'humain dans la trame #894 et le véhicule dans la trame #1115 sur les n_f = 15 trames consécutives suivant leur première apparition dans la scène.

5.4.3. Séquence 3 : Changements d'apparence

Le principal défi de cette séquence est la présence, dans la scène, d'un humain en mouvement subissant de fortes variations d'apparence. Quelques trames représentatives de cette séquence montrant la comparaison des résultats de notre méthode de suivi proposée avec ceux des méthodes ASLA, L1-APG, TGPR, CSK, PF+Int et PF+Tex sont illustrées dans la Figure 5.16. Notons ici que, comme notre méthode proposée a automatiquement initialisé le suivi de l'humain à la trame #46, et pour une comparaison équitable, nous avons initialisé tous les autres trackers manuellement en utilisant la même boîte englobante d'initialisation que

celle de notre méthode proposée. Ensuite, chaque tracker continue à suivre l'humain automatiquement jusqu'à ce qu'il disparaisse de la scène.

D'après les résultats de la Figure 5.16, nous pouvons observer qu'à partir de la trame #46 jusqu'à environ la trame #357, tous les trackers performent plutôt bien, excepté le tracker PF+Tex qui semble dériver légèrement lorsque l'humain passe à travers des régions d'arrière-plan ayant des caractéristiques de texture similaires à celles de cet humain. Cela peut être observé au environ des trames #65 et #277, lorsque l'humain passe devant les lampadaires.

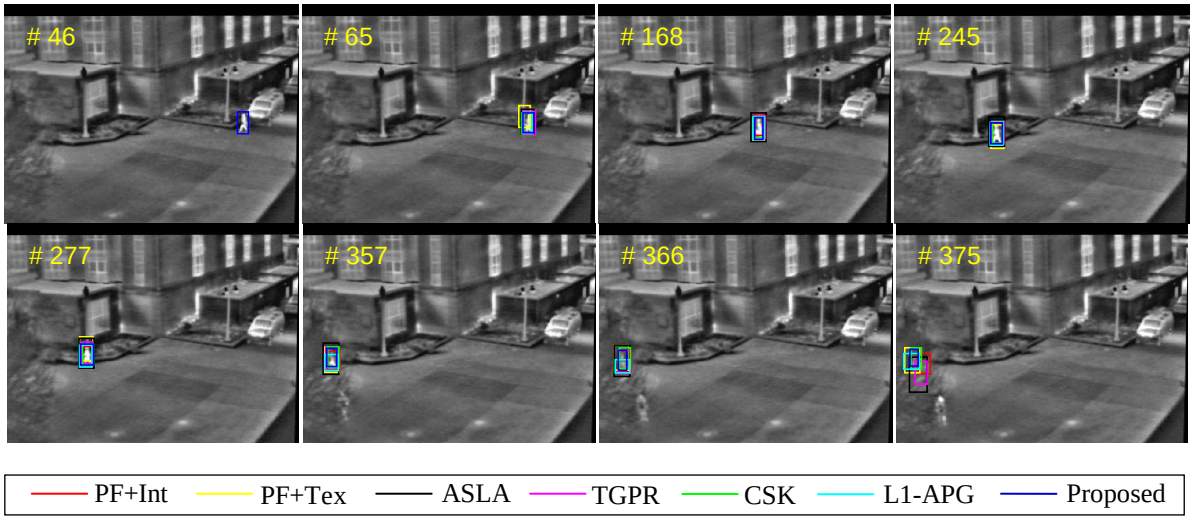
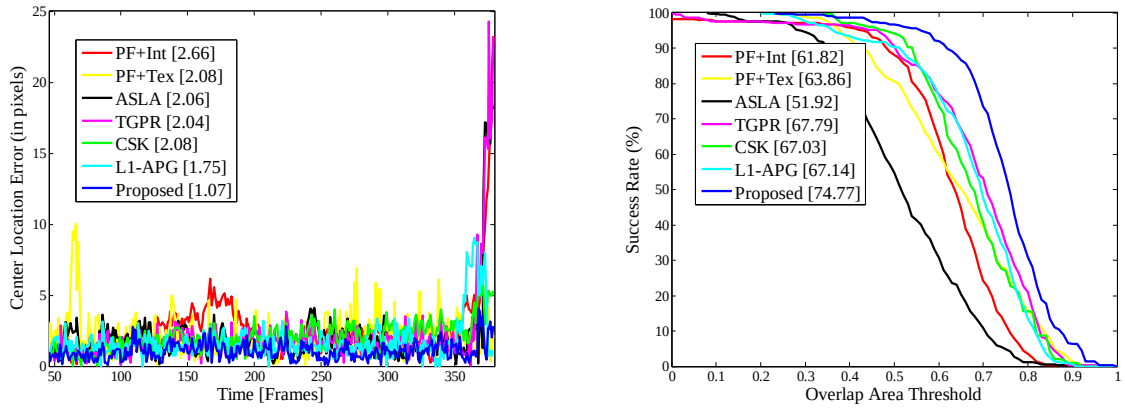


Figure 5.16: Résultats expérimentaux des différents trackers sur la Séquence 3 (changements sévères d'apparance).

Cependant, à partir de la trame #357, l'humain suivi commence à se déplacer derrière les branches d'arbres, et ceci provoque un changement drastique dans l'intensité de son apparence. Ainsi, le principal défi pour tous les trackers est de garder la piste de l'humain après que ces changements d'apparence aient eu lieu. D'après les résultats de la Figure 5.16, nous pouvons observer que, parmi tous les trackers comparés, seules les méthodes PF+Tex, CSK et notre méthode peuvent encore garder la piste de l'humain après les changements d'apparence, tandis que les autres trackers, c'est-à-dire les méthodes PF+Int, ASLA et TGPR, se sont complètement dérivées de l'humain suivi. Quant aux résultats du tracker L1-APG, ils sont moins satisfaisants, car le tracker continue à suivre l'humain, mais avec une taille inappropriée de la boîte englobante.

Les résultats de comparaisons quantitatives entre les différentes méthodes en termes de CLE et de SR sont présentés dans la Figure 5.17. Les valeurs indiquées dans la légende des graphiques CLE sont les erreurs moyennes de localisation du centre et celles indiquées dans la légende des graphiques SR sont les scores de l'aire sous la courbe (Area Under the Curve, AUC). À partir des différents tracés, nous pouvons observer que, parmi tous les trackers comparés, notre méthode proposée fournit la meilleure performance en termes de CLE et de SR, avec une erreur moyenne de 1.07 pixels et un score AUC de 74.77%. Cependant, sur cette séquence, le tracker PF+Int présente la plus faible performance en termes de CLE, tandis que le tracker ASLA présente la plus faible performance en termes de SR.



(a) (b)

Figure 5.17: Comparaison entre les différents trackers en termes de (a) CLE, et (b) SR sur la Séquence 3.

Les résultats de comparaison entre les différentes méthodes en termes de FPS moyen (vitesse de suivi moyenne) sont présentés dans le Tableau 5.6. D'après ce tableau, nous pouvons observer que, outre ses meilleures performances, notre méthode de suivi proposée peut fonctionner à environ 15 FPS en moyenne, ce qui est environ 3 fois plus rapide que les trackers ASLA et L1-APG, et 31 fois plus rapide que le tracker TGPR. Cependant, en raison de la combinaison de plusieurs caractéristiques, notre méthode proposée prend plus de temps que le tracker CSK, qui utilise uniquement la caractéristique d'intensité.

Tableau 5.6: FPS moyen pour différents trackers.

Méthode	**FPS moyen**
ASLA	05.80
L1-APG	05.14
TGPR	0.47
CSK	170.96
PF+Int	15.77
PF+Tex	09.47
Proposée	14.84

Afin de démontrer les avantages et l'efficacité de notre procédure proposée pour l'ajustement automatique des poids des caractéristiques (Section 3.4.3.5 du Chapitre 3), nous montrons dans la Figure 5.18 les tracés de la variation trame par trame des poids $\widehat{\omega}_{k,1}, \widehat{\omega}_{k,2}$ et $\widehat{\omega}_{k,3}$ assignés, respectivement, aux caractéristiques d'intensité, de texture RLBP et de vélocité de mouvement pendant le suivi de l'humain dans la Séquence 3. D'après ces tracés, nous pouvons observer qu'avant les changements d'apparence, les caractéristiques d'intensité et de texture RLBP sont plus fiables que la vélocité de mouvement. Cependant, lorsque les changements d'apparence apparaissent à partir d'environ la trame #355, la caractéristique d'intensité devient moins fiable et moins discriminante que les autres caractéristiques. En conséquence, son poids correspondant est automatiquement diminué par notre méthode, tandis que les poids des caractéristiques de texture RLBP et de vélocité de mouvement sont augmentés pour continuer à suivre l'humain de manière plus robuste.

Dans la Figure 5.19, nous montrons également la variation automatique des valeurs des taux $\alpha_{k,Int}$ et $\alpha_{k,RLBP}$ pour mettre à jour, respectivement, les modèles d'intensité et de texture RLBP pendant le suivi de l'humain dans la Séquence 3. Sur cette figure, nous pouvons observer qu'avant l'apparition des changements d'apparence, notre méthode proposée adapte de façon régulière les modèles d'intensité et de texture RLBP avec des valeurs de taux fluctuant autour d'une valeur constante sans aucune variation significative. Cependant, lorsque les changements d'apparence apparaissent à partir de la trame #355, nous pouvons observer que notre méthode proposée peut s'adapter de manière robuste à ces changements en ralentissant le processus de mise à jour du modèle d'intensité.

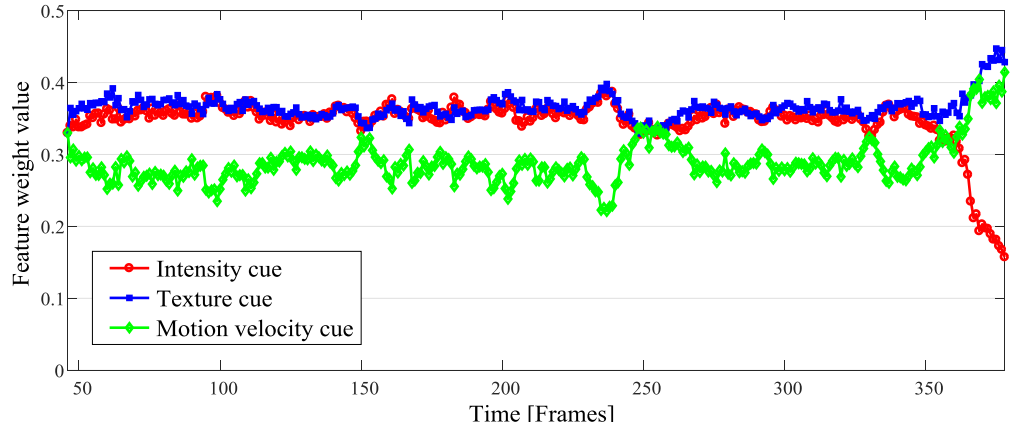


Figure 5.18: Variation des valeurs des poids $\hat{\omega}_{k,1}$, $\hat{\omega}_{k,2}$ et $\hat{\omega}_{k,3}$ attribuées aux caractéristiques d'intensité, de texture RLBP et de vélocité de mouvement pendant le suivi de l'humain dans la Séquence 3.

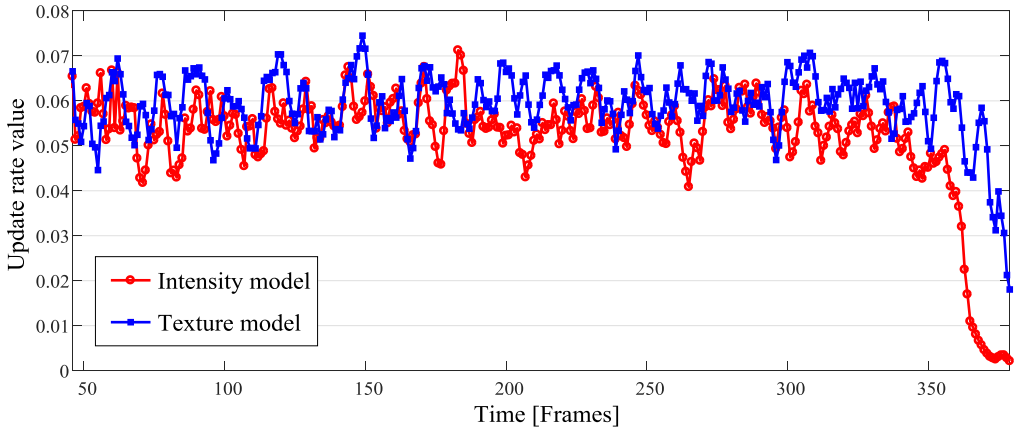


Figure 5.19: Variation des valeurs des taux $\alpha_{k,\mathrm{Int}}$ et $\alpha_{k,\mathrm{RLBP}}$ pour la mise à jour des modèles d'intensité et de texture RLBP.

5.4.4. Séquence 4 : Occultation inter-humain

Cette séquence contient deux humains en mouvement dans le champ de vision de la caméra. Le premier entre dans la scène par le côté droit et la quitte par le côté gauche, tandis que le second entre dans la scène par le côté gauche et la quitte par le côté droit. Ainsi, le principal défi pour tous les trackers dans cette séquence est la présence d'une occultation inter-humain au milieu de la scène. Quelques trames représentatives de cette séquence montrant les résultats des différents trackers lorsqu'ils sont appliqués pour le suivi du premier et du second humain sont illustrées dans les Figures 5.20(a) et 5.20(b), respectivement.

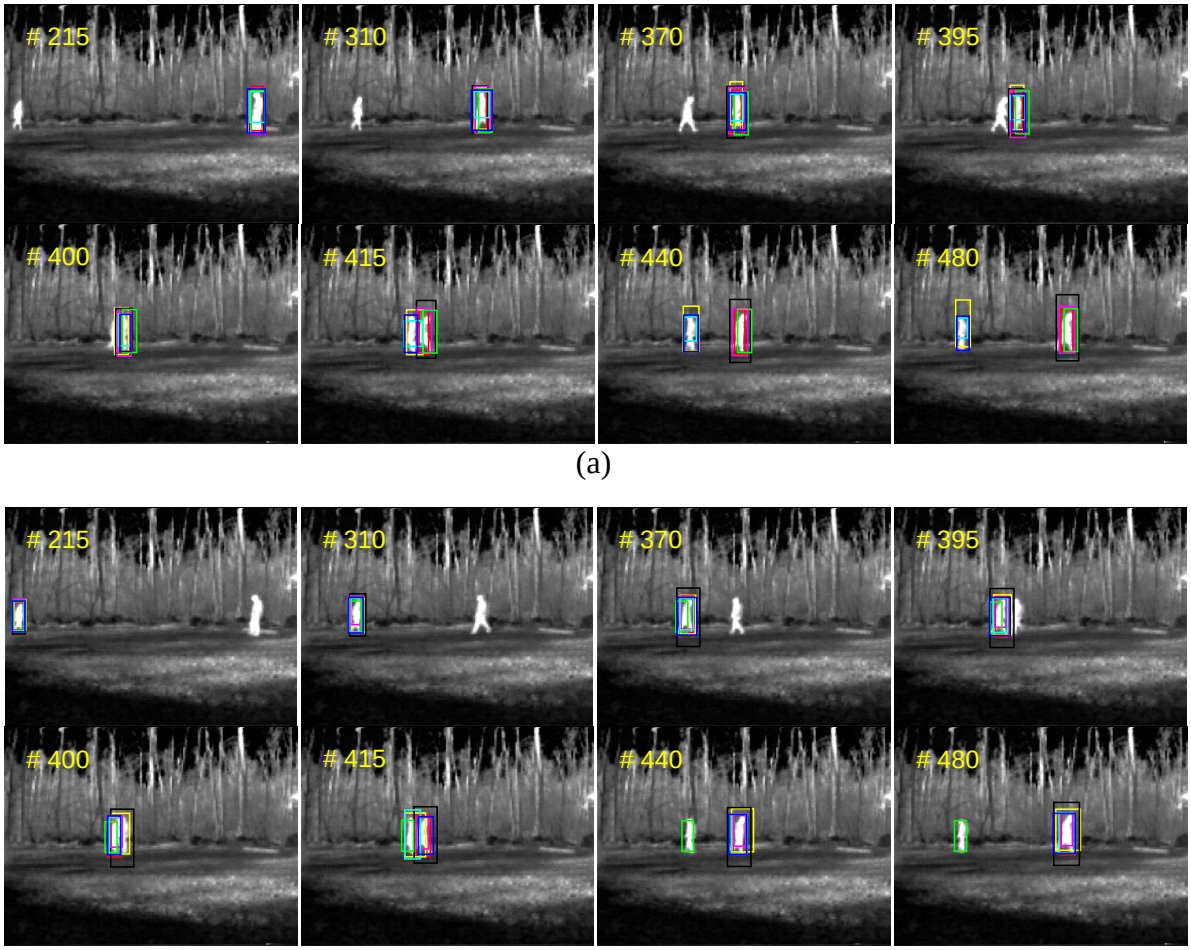
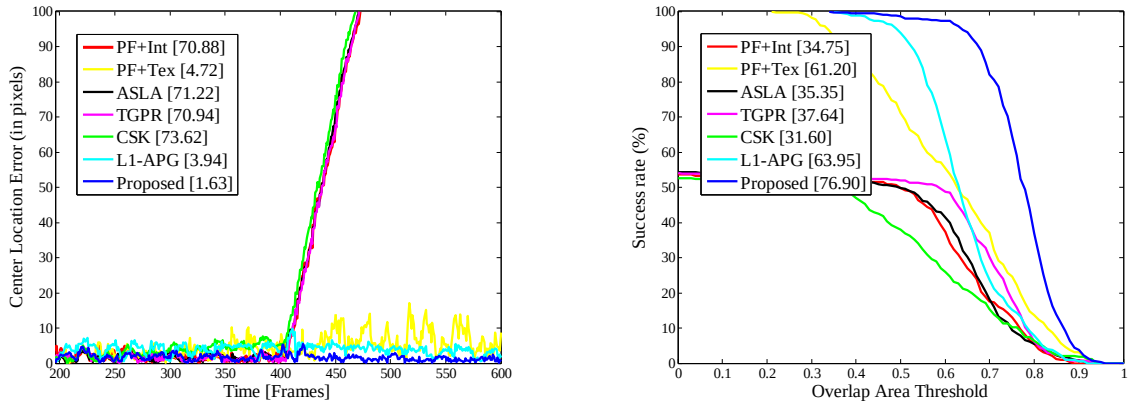


Figure 5.20: Résultats expérimentaux de différents trackers sur la Séquence 4 (occultation inter-humain). (a) Résultats du suivi du premier humain. (b) Résultats du suivi du deuxième humain.

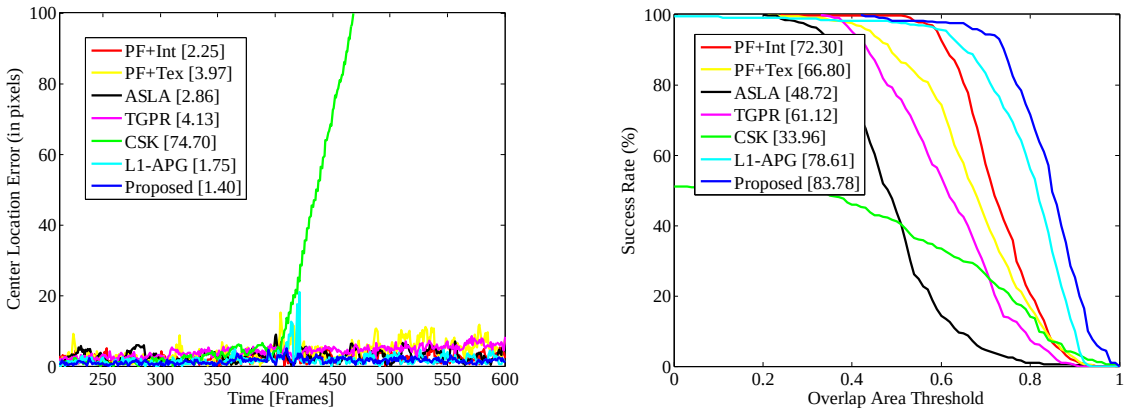
Sur la Figure 5.20(a), nous pouvons observer qu'avant l'apparition de l'occultation inter-humain, tous les trackers peuvent suivre avec succès le premier humain, malgré quelques différences de précision entre eux. Cependant, après l'apparition de l'occultation inter-humain aux alentours de la trame #400, seules les méthodes PF+Tex, L1-APG et notre méthode proposée peuvent suivre correctement le premier humain, tandis que les autres trackers échouent et passent au suivi du deuxième humain. D'après les résultats de la Figure 5.20(b), nous pouvons observer que tous les trackers, à l'exception de CSK, peuvent maintenir le suivi du second humain jusqu'à la fin de la séquence. En observant les trames de la Figure 5.20, et plus

particulièrement la trame #400 et la trame #415, nous pouvons également voir que notre méthode proposée peut détecter avec succès l'occultation inter-humain et ce, en effectuant le processus de séparation entre les deux humains.



(a) (b)

Figure 5.21: Comparaison en termes de (a) CLE, et (b) SR entre les différents trackers pour le suivi du premier humain dans la Séquence 4.



(a) (b)

Figure 5.22: Comparaison en termes de (a) CLE, et (b) SR entre les différents trackers pour le suivi du deuxième humain dans la Séquence 4.

Dans les Figures 5.21 et 5.22, nous montrons les résultats de comparaison quantitative entre les différents trackers en termes de CLE et SR. D'après les différents graphiques, nous pouvons observer que, parmi tous les trackers comparés, notre méthode proposée obtient les meilleures performances sur l'ensemble de la séquence, en fournissant une CLE moyenne plus faible et une aire

sous la courbe SR plus élevée. La deuxième meilleure performance sur cette séquence est obtenue par le tracker L1-APG.

5.4.5. Séquence 5 : Occultation humain/arrière-plan

Cette séquence contient un humain entrant dans la scène par le côté gauche de la caméra, se déplaçant horizontalement vers le côté droit jusqu'à ce qu'il disparaisse complètement de la scène. Au cours de son déplacement, et en raison d'une occultation par des objets d'arrière-plan (arbres), l'humain disparaît temporairement du champ de vision de la caméra dans certaines trames, puis réapparaît dans les trames subséquentes. Le principal défi consiste donc à suivre l'humain en mouvement sur l'ensemble de la séquence, malgré la présence de l'occultation humain/arrière-plan. Quelques trames représentatives de cette séquence montrant les résultats des différents trackers sont présentées dans la Figure 5.23. Sur cette figure, nous pouvons observer que tous les trackers peuvent suivre avec succès l'humain jusqu'à la trame #427. Cependant, après l'apparition de l'occultation humain/arrière-plan aux alentours de la trame #434, nous pouvons observer que seules les méthodes PF+Int, ASLA, L1-APG et notre méthode proposée sont capables de recapturer l'humain suivi, alors que les trackers PF+Tex, TGPR et CSK perdent complètement l'humain et ne le récupèrent jamais dans les trames ultérieures.

Les résultats de comparaison quantitative entre les différents trackers en termes de CLE et SR sont présentés dans les Figures 5.24(a) et 5.24(b), respectivement. D'après ces figures, nous pouvons observer que les courbes CLE des trackers PF+Tex, TGPR et CSK divergent totalement après l'apparition de l'occultation humain/arrière-plan aux alentours de la trame #434. Les autres trackers peuvent suivre avec précision l'humain sur toute la séquence, à l'exception du tracker ASLA qui produit une estimation inappropriée de la taille de l'humain suivi, fournissant ainsi les mauvaises performances en termes d'aire sous la courbe SR.

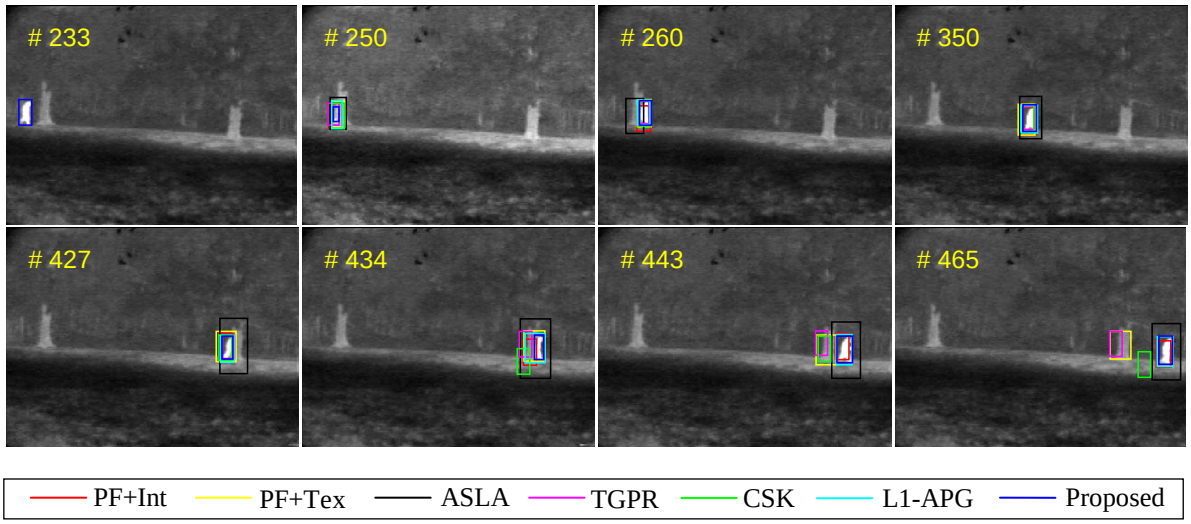
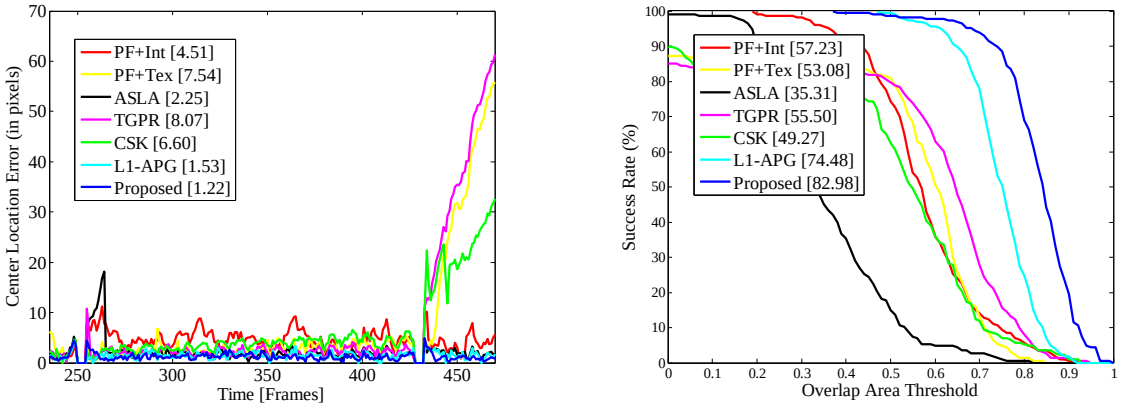


Figure 5.23: Résultats expérimentaux des différents trackers sur la Séquence 5 (occultation humain/arrière-plan).



(a) (b)

Figure 5.24: Comparaison en termes de (a) CLE, et (b) SR entre les différents trackers sur la Séquence 5.

5.4.6. Séquence 6 : Changements d'échelle

Cette séquence contient un humain tenant une arme à feu se déplaçant de loin vers la position de la caméra. Ce mouvement entraîne un changement significatif de la taille (échelle) de l'humain au cours du temps. Ainsi, le principal défi pour tous les trackers dans cette séquence est de s'adapter à ces changements de taille tout au long du processus de suivi. La Figure 5.25 présente quelques trames de cette séquence montrant les résultats des différents trackers. Sur cette figure, nous pouvons observer que, lorsque l'humain est en mouvement vers la

caméra, seules les méthodes ASLA, L1-APG et notre méthode proposée peuvent s'adapter efficacement aux changements d'échelle. Le reste des trackers, à savoir PF+Int, PF+Tex, TGPR et CSK, sont incapables de gérer ces changements d'échelle, car la taille de leurs boîtes englobantes estimées est trop petite par rapport à la taille réelle de l'humain suivi.

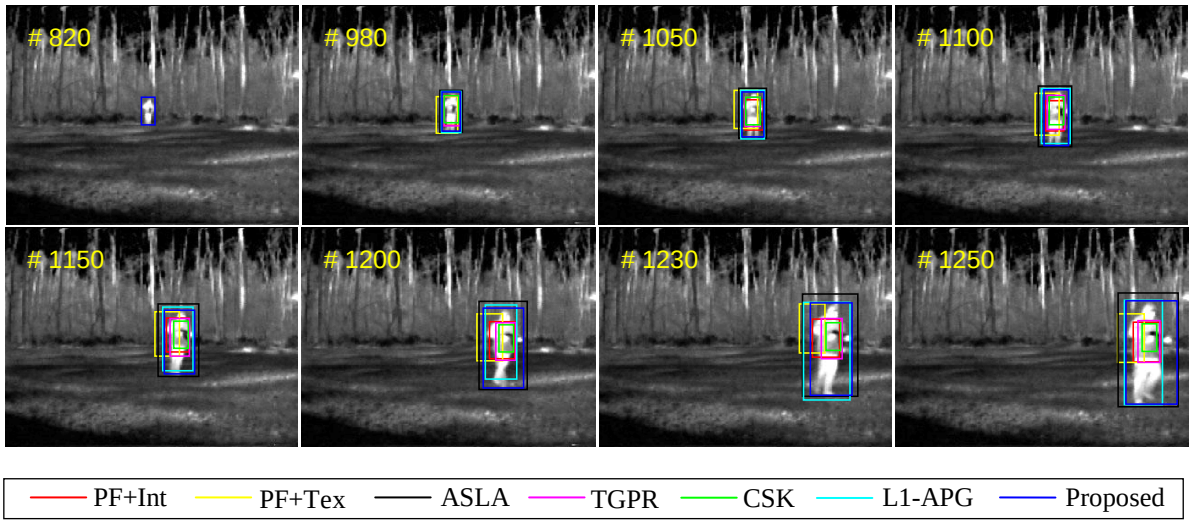
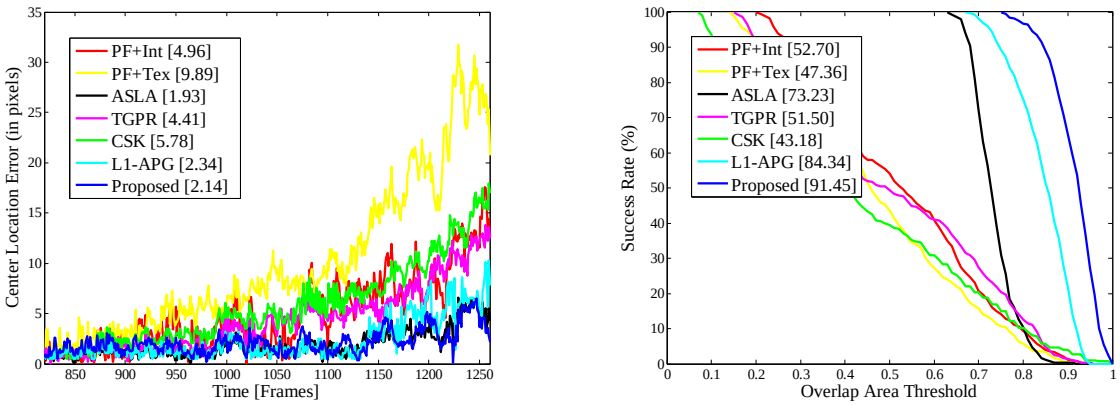


Figure 5.25: Résultats expérimentaux des différents trackers sur la Séquence 6 (changements d'échelle).



(a) (b)

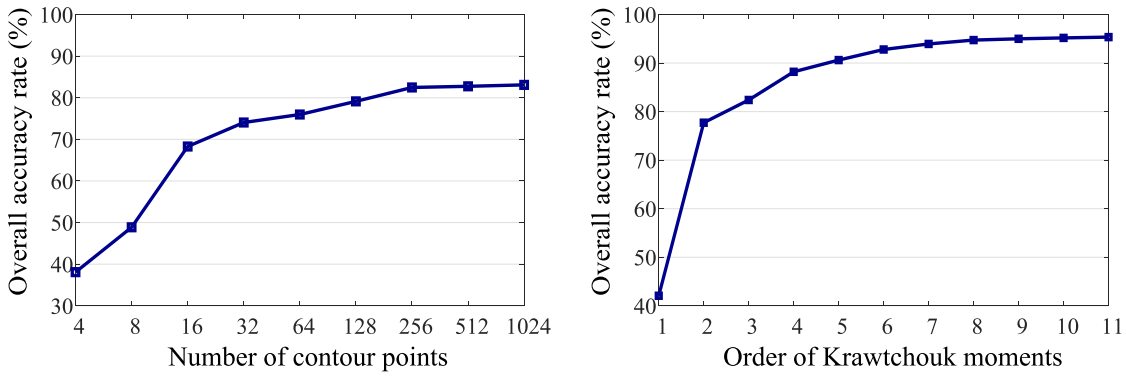
Figure 5.26: Comparaison en termes de (a) CLE, et (b) SR entre les différents trackers sur la Séquence 6.

Les résultats de comparaison quantitative entre les différents trackers en termes de CLE et de SR sont présentés dans les Figures 5.26(a) et 5.26(b), respectivement. En observant les graphiques, nous pouvons clairement observer que notre méthode proposée atteint les meilleures performances en termes d'aire sous la courbe SR, fournissant un score de 91.45%. Ces résultats démontrent la grande capacité de notre méthode à faire face à de grands changements d'échelle de l'humain suivi.

5.5. Evaluation des approches de reconnaissance de posture

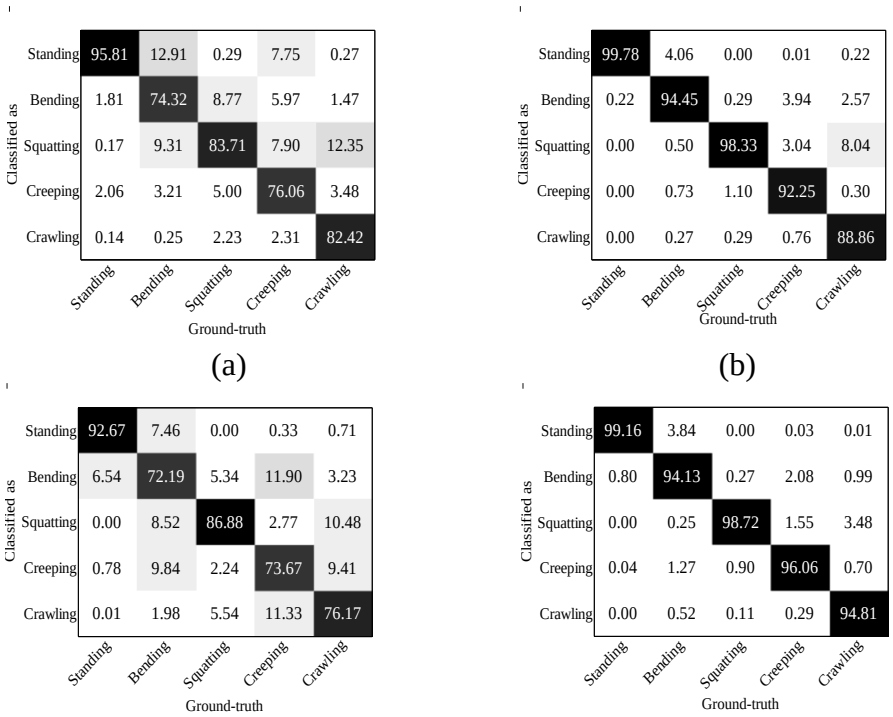
Afin d'évaluer efficacement les performances de notre approche de reconnaissance de postures proposée, nous avons mené une procédure de validation croisée à 10 blocs (folds) dans laquelle la totalité des images de notre base de données de postures créée sont divisées aléatoirement en 10 groupes contenant le même nombre d'échantillons. Ensuite, à chaque essai de la validation croisée, un des 10 groupes est utilisé pour le test, tandis que les autres sont laissés pour l'apprentissage. La procédure de validation croisée est arrêtée lorsque tous les groupes sont utilisés pour le test. La précision globale est donc estimée comme la moyenne des précisions obtenues à partir des 10 essais réalisés. Afin d'éviter une évaluation biaisée en raison du caractère aléatoire du processus d'échantillonnage, nous avons également répété la validation croisée 50 fois et la performance finale est obtenue en calculant la moyenne des performances de l'ensemble des 50 expériences de validation croisée.

Cependant, comme les trois types de caractéristiques (CCH, moments de Krawtchouk, et les caractéristiques géométriques) décrites dans la Section 4.3.1 caractérisent différents aspects de la silhouette du corps humain, nous les évaluons d'abord indépendamment, puis en combinaison, afin d'étudier leurs effets séparés et combinés sur la performance globale de la reconnaissance de posture.



(a) (b)

Figure 5.27: Résultats de la reconnaissance de posture en utilisant (a) CCH avec différentes valeurs de points de contour, et (b) les moments de Krawtchouk avec différents ordres.



(a) (b)

(c) (d)

Figure 5.28: Matrices de confusion (en %) des résultats de reconnaissance de posture obtenus en utilisant (a) le descripteur CCH (avec N_p = 256), (b) les moments de Krawtchouk de 8ème ordre, (c) les caractéristiques géométriques, et (d) toutes les caractéristiques combinées.

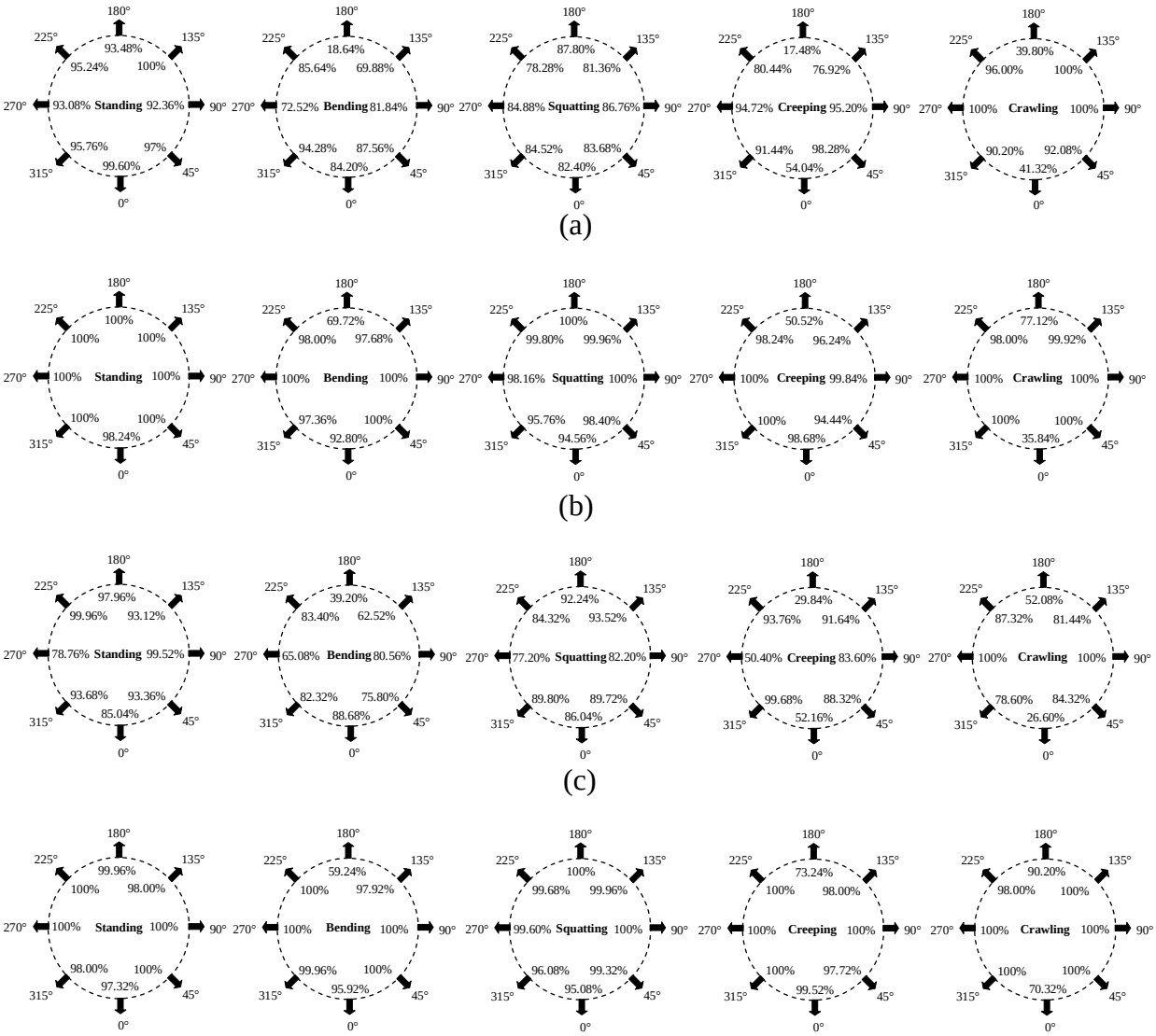


Figure 5.29: Précisions détaillées pour chaque posture et direction de mouvement obtenues par (a) le CCH (avec N_p = 256), (b) les moments de Krawtchouk de 8ème ordre, (c) les caractéristiques géométriques, et (d) toutes les caractéristiques combinées.

5.5.1. Histogramme de chaine de codes

Pour trouver la meilleure valeur pour le nombre de points N_p inclus dans le contour ré-échantillonné (Figure 4.1 du Chapitre 4), nous avons mené plusieurs expériences dans lesquelles nous avons testé plusieurs valeurs pour N_p à partir de l'ensemble {4, 8, 16, 32, 64, 128, 256, 512 et 1024}. Les résultats obtenus pour chaque valeur de N_p sont présentés dans la Figure 5.27(a). D'après cette courbe,

nous pouvons observer que l'augmentation de la valeur de N_p jusqu'à 256 améliore considérablement les performances de la reconnaissance de posture, mais au-delà de cette valeur, les performances n'évoluent que très peu. Ainsi, sur la base de ces résultats, nous pouvons conclure que le ré-échantillonnage du contour de la silhouette humaine à 256 points est suffisant pour le problème de la reconnaissance de posture humaine en utilisant le descripteur CCH, produisant une précision globale d'environ 82.46%.

Afin de décrire plus en détail les résultats du descripteur CCH (avec le paramètre N_p =256) et de comprendre quel type de posture humaine et de direction de mouvement provoque le plus d'erreurs de classification, nous montrons dans la Figure 5.28(a) et la Figure 5.29(a), respectivement, la matrice de confusion et les précisions détaillées obtenues pour chaque type de posture et direction de mouvement contenu dans notre base de données créée. À partir de la matrice de confusion, nous pouvons observer que parmi les cinq types de posture, la posture "Standing" est la plus susceptible d'être correctement reconnue par le descripteur CCH, alors que les postures "Bending" et "Creeping" sont les plus difficiles à reconnaître. Ce résultat peut être expliqué par le fait que la posture "Standing" est la moins affectée par les changements de direction de mouvement, comme nous pouvons le voir sur la Figure 5.29(a), contrairement aux postures "Bending" et "Creeping", qui semblent être difficiles à reconnaître lorsqu'elles sont observées dans certaines directions, en particulier à 180°. D'après les précisions détaillées sur la Figure 5.29(a), nous pouvons également constater que certaines directions de mouvement sont plus discriminantes pour certaines postures. Par exemple, la posture "Crawling" est parfaitement reconnue lorsqu'elle est observée à 90° ou 270°, mais elle est souvent erronément classée dans une autre posture lorsqu'elle est observée à 0° ou 180°.

5.5.2. Moments de Krawtchouk

Pour déterminer la valeur appropriée de l'ordre $(n+m)$ des moments de Krawtchouk (Section 4.3.1.1 du Chapitre 4), nous avons mené des expériences extensives dans lesquelles plusieurs valeurs d'ordre, allant de 3 à 13, ont été testées. Les résultats obtenus pour chaque valeur d'ordre sont présentés dans la Figure 5.27(b). Sur cette figure, nous pouvons observer que l'augmentation de l'ordre des moments de Krawtchouk jusqu'à 8 améliore progressivement la précision

de reconnaissance de posture, mais l'augmentation de cet ordre au-delà de 8 n'améliore pas substantiellement les performances de ces moments. Ainsi, sur la base de ces résultats, nous pouvons conclure que l'utilisation des moments de Krawtchouk d'ordre 8 est suffisante pour le problème de reconnaissance de posture humaine, fournissant une précision globale d'environ 94.73%.
La matrice de confusion et les précisions détaillées pour chaque type de posture humaine et chaque direction de mouvement obtenues en utilisant les moments de Krawtchouk de 8ème ordre sont présentées dans la Figure 5.28(b) et la Figure 5.29(b), respectivement. À partir de la matrice de confusion, nous pouvons observer que, comme pour le descripteur CCH, la posture "Standing" est la plus correctement reconnue par les moments de Krawtchouk, tandis que la posture "Crawling" est la plus confondue avec d'autres postures. Les précisions détaillées de la Figure 5.29(b) montrent que les moments de Krawtchouk sont moins sensibles aux changements de direction de mouvement par rapport au descripteur CCH. Cependant, la reconnaissance efficace de certaines postures, telles que "Bending" et "Creeping" dans la direction 180° reste encore difficile à atteindre. Cela est dû au fait que, lorsqu'elles sont observées dans cette direction, les silhouettes correspondantes à ces postures apparaissent visuellement très similaires les unes aux autres, avec des caractéristiques de forme moins discriminantes (Figure 5.3).

5.5.3. Caractéristiques géométriques

La matrice de confusion obtenue en utilisant les caractéristiques géométriques est présentée dans la Figure 5.28(c), et les précisions détaillées obtenues pour chaque type de posture humaine et direction de mouvement sont présentées dans la Figure 5.29(c). D'après ces figures, nous pouvons observer que, par rapport aux moments de Krawtchouk, les caractéristiques géométriques seules offrent moins de performances en termes de précision de reconnaissance de posture. Cependant, leurs performances dans l'ensemble sont très proches de ceux du descripteur CCH. D'après la matrice de confusion de la Figure 5.28(c), nous pouvons observer que les postures "Standing" et "Squatting" sont les plus discriminées par les caractéristiques géométriques, alors que les postures "Bending" et "Creeping" sont les plus difficiles à reconnaître. Ces résultats peuvent être expliqués par le fait que les caractéristiques géométriques sont globales et elles ne permettent pas de capturer les détails locaux de certains types de postures

relativement complexes telles que "Bending" et "Creeping". Cela peut être également expliqué par les résultats de la Figure 5.29(c), qui montrent une robustesse relativement faible des caractéristiques géométriques face aux changements de direction de mouvement, en particulier lorsque l'humain est en mouvement dans la posture "Bending" ou "Creeping".

5.5.4. Combinaison des caractéristiques

Après avoir évalué indépendamment le descripteur CCH, les moments de Krawtchouk et les caractéristiques géométriques, nous montrons dans la Figure 5.28(d) la matrice de confusion de la reconnaissance de posture obtenue en combinant toutes les caractéristiques. Dans la Figure 5.29(d), nous montrons également les précisions détaillées obtenues pour chaque type de posture humaine et direction de mouvement. D'après la matrice de confusion, nous pouvons observer que la combinaison de toutes les caractéristiques améliore la précision de reconnaissance d'environ 3.81% pour la posture "Creeping" et d'environ 5.95% pour la posture "Crawling", par rapport aux meilleurs résultats obtenus par les moments de Krawtchouk seuls. D'après les précisions détaillées de la Figure 5.29(d), nous pouvons observer que, pour certaines directions de mouvement, la combinaison des caractéristiques est considérablement plus performante que le meilleur résultat obtenu par une seule caractéristique. Par exemple, dans le cas de la reconnaissance de la posture "Creeping" dans la direction 180°, la combinaison améliore le taux de précision d'environ 22.72% par rapport au meilleur résultat obtenu par les moments de Krawtchouk, et dans le cas de la reconnaissance de la posture "Crawling" dans la direction 0°, la combinaison améliore le taux de précision d'environ 29% par rapport au meilleur résultat obtenu par le descripteur CCH. Ces résultats signifient que lorsqu'elles sont combinées, les différentes caractéristiques sont complémentaires les unes des autres et fournissent une représentation plus précise et plus fiable de la posture humaine.

Afin de démontrer davantage l'efficacité de notre approche de reconnaissance de posture humaine proposée basée sur la combinaison de caractéristiques, nous avons comparé nos résultats avec ceux obtenus en utilisant d'autres caractéristiques d'état de l'art, à savoir les histogrammes de projection horizontale et verticale (I. Haritaoglu et al., 2000; Goldmann et al., 2004; Cucchiara et al., 2005; Boulay et al., 2006; Miao Yu et al., 2012), les moments de Hu (Boulay et al.,

2006; Feng and Lin, 2010), le squelette étoile (Collins et al., 2000; Boulay et al., 2006; Chen et al., 2006), et les caractéristiques profondes (Adhikari et al., 2017). Les résultats de comparaison sont présentés dans la Figure 5.30. D'après la figure, nous pouvons observer que, dans l'ensemble, notre approche proposée fournit les meilleures performances de reconnaissance de posture. Plus précisément, elle surpasse les histogrammes de projection horizontale et verticale par 2.6 %, les moments de Hu par 7 %, et le squelette étoile par 8.9 %. Cette dernière est la caractéristique la moins performante sur notre base de données. Quant aux caractéristiques profondes, elles donnent des performances très proches de celles de notre méthode, avec seulement 0.53 % de différence en termes de précision globale par rapport à notre méthode. Cette légère sous-performance est due principalement au fait que les caractéristiques profondes sont très gourmandes en données durant l'apprentissage, et comme notre base de données de postures contient uniquement 2000 images, cela peut être considéré insuffisant pour entraîner un réseau de neurones convolutifs afin qu'il puisse atteindre ses performances optimales. Dans ce cas, notre méthode basée sur des caractéristiques multiples s'avère être plus efficace, en plus d'être moins coûteuse en ressources de calcul, contrairement aux caractéristiques profondes, qui nécessitent une plus grande complexité de calcul, notamment pendant la phase d'apprentissage.

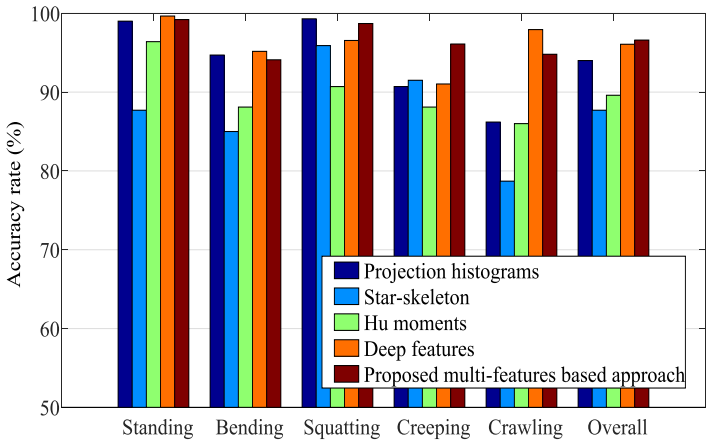


Figure 5.30: Comparaison des résultats de notre approche avec ceux obtenus par des caractéristiques d'état de l'art.

5.6. Application : détection d'un comportement humain anormal

Dans cette section, nous présentons un exemple d'application de notre système proposé pour la détection d'un comportement humain anormal, qui est "la chute" (falling). Cet événement anormal est l'un des risques de santé les plus courants et les plus graves auxquels les gens sont exposés dans leur vie quotidienne, en particulier dans des conditions de mauvaise visibilité, telles que la nuit. Lorsqu'il se produit, cet incident peut être très dangereux pour la personne tombant, car cela peut lui causer de graves blessures physiques et psychologiques, notamment lorsqu'il s'agit de personnes âgées ou de personnes ayant des besoins particuliers. La nécessité d'un système intelligent capable d'alerter automatiquement le personnel du foyer, ou du centre hospitalier, pour qu'il puisse intervenir le plus rapidement possible est donc primordiale.

La séquence de test capturée contient un nombre total de 960 trames et représente un humain descendant des escaliers, lorsqu'il perd soudainement l'équilibre puis tombe, reste au sol pendant un moment, puis se lève lentement et continue à marcher pour quitter la scène. Quelques trames représentatives de cette séquence avec le résultat de notre système d'analyse automatique du comportement humain sont présentées dans la Figure 5.31. Dans cette figure, la courbe jaune représente la trajectoire estimée de l'humain suivi, le rectangle vert représente sa boîte englobante, tandis que le rectangle bleu représente la zone surveillée (préalablement définie par l'utilisateur), et son objectif est de réduire davantage les fausses alarmes causées par les mouvements (clutters) de l'environnement. Notons ici, qu'afin de distinguer un comportement anormal d'un comportement normal, nous avons étiqueté un humain se déplaçant dans la posture "Standing" comme normal, tandis qu'un humain se déplaçant dans l'une des autres postures, c.-à-d., "Bending", "Squatting", "Creeping", et "Crawling", comme anormal. D'après les résultats de la Figure 5.31, nous pouvons observer que, lorsque l'humain suivi tombe sur le sol, un changement de sa posture corporelle de "Standing" vers "Crawling" se produit. Et, comme cette situation dure pour un temps suffisamment long (dans notre cas, plus de 3s), notre système proposé passe de l'état "No abnormal behaviour" (pas de comportement anormal) vers l'état "Abnormal behaviour detected" (comportement anormal détecté), et ce afin d'informer l'utilisateur que des actions préventives appropriées doivent être prises.

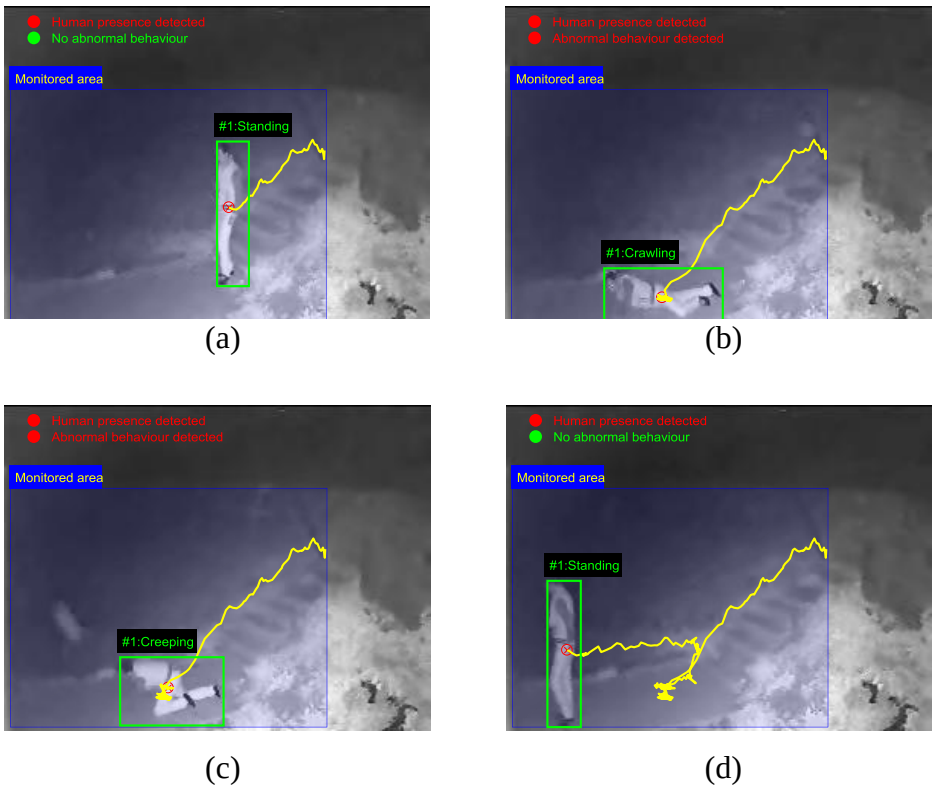
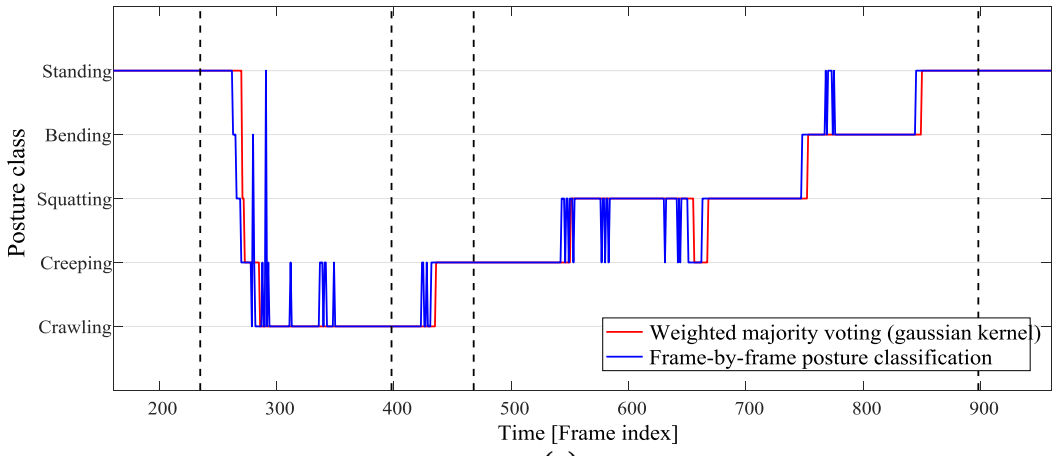
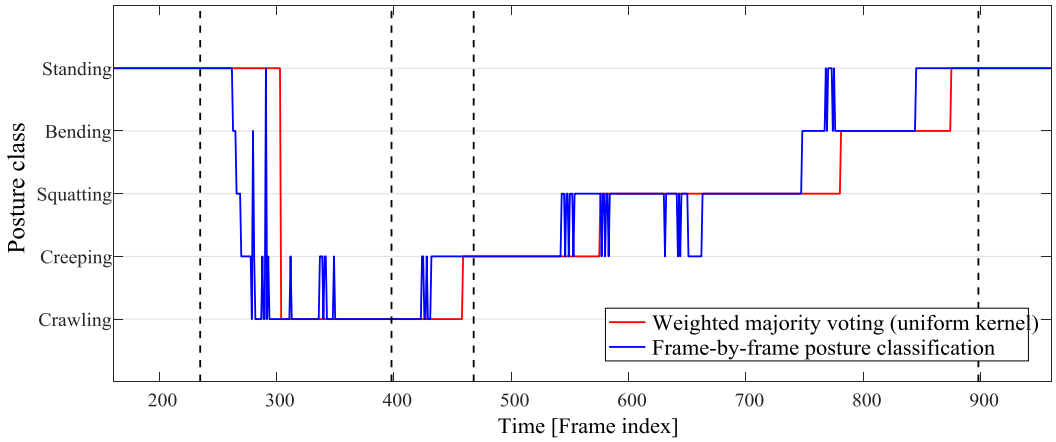


Figure 5.31 : Quelques trames de la séquence "chute" avec le résultat de notre système d'analyse automatique du comportement humain. (a) Trame #235. (b) Trame #398. (c) Trame #468. (d) Trame #898.

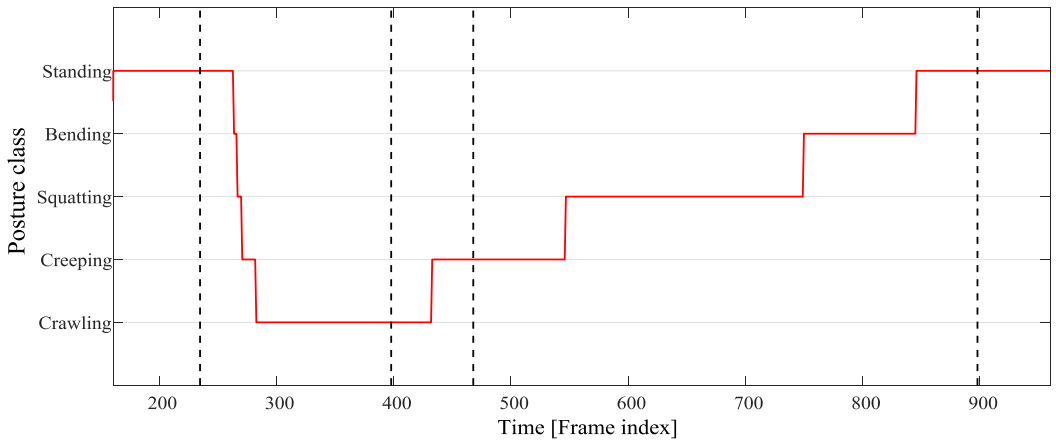
Afin de mieux comprendre les résultats montrés dans les Figures 5.31(a-d), nous montrons sur la Figure 5.32(a) la courbe d'analyse du comportement obtenue en utilisant le filtre WMV (avec un noyau gaussien, et une longueur de l'historique temporel $T = 60$), et la courbe obtenue sans le filtrage temporel de posture (c.-à-d., par classification trame par trame). En comparant ces courbes avec la vérité terrain présentée dans la Figure 5.32(c), nous pouvons constater que sans le filtrage de posture, le résultat de l'analyse du comportement humain présente un certain nombre d'erreurs de classification, qui sont principalement dues à des silhouettes humaines mal segmentées, ainsi qu'à des confusions entre certains types de postures. Cependant, la comparaison montre que l'application du filtre WMV avec un noyau gaussien réduit considérablement le nombre d'erreurs de classification de posture et améliore significativement le résultat global de l'analyse du comportement humain.



(a)



(b)



(c)

Figure 5.32 : Résultat d'analyse du comportement sur la séquence "chute". (a) Résultat en utilisant le filtre WMV avec un noyau gaussien et une longueur de l'historique temporel $T = 60$. (b) Résultat en utilisant le filtre WMV sans le noyau gaussien (avec un noyau uniforme). (c) Vérité-terrain.

Pour des fins de comparaison, nous montrons sur la Figure 5.32(b) la courbe obtenue en utilisant le filtre WMV mais sans le noyau gaussien (c.-à-d., en utilisant des poids uniformes). En comparant cette courbe avec la vérité terrain montrée dans la Figure 5.32(c), nous pouvons observer que l'application du filtre WMV sans le noyau gaussien réduit significativement le taux d'erreurs de classification de posture. Cependant, un retard important (d'environ 30 trames) est introduit à la sortie du filtre WMV, ce qui peut être non-négligeable notamment lorsque nous avons à faire face à un événement très dangereux comme la chute.

5.7. Conclusion

Dans ce chapitre, nous avons présenté les résultats expérimentaux réalisés afin d'évaluer les différentes étapes constituant notre système de vidéo surveillance intelligent proposé. Dans une première partie, nous avons décrit les différentes bases de données utilisées dans le cadre de ce projet, dont certaines sont des bases de données publiques, et d'autres ont été créées, par nous-mêmes, spécialement pour ce projet. Ensuite, dans une deuxième partie, nous avons présenté les résultats de détection de personnes obtenus, et qui ont montré la capacité de nos deux approches proposées à discriminer un être humain de toute autre forme d'objet non humain en mouvement (véhicules, animaux, etc.), et ce même en cas de présence de changements de posture et de direction de mouvement de ces humains. Dans une troisième partie, nous avons présenté les expériences réalisées pour évaluer l'approche de suivi proposée, qui est basée sur un filtre à particules et une combinaison adaptative de multiples caractéristiques. L'étude comparative avec certaines approches reportées dans la littérature a montré l'efficacité et la supériorité de notre approche proposée pour le suivi de personnes dans des environnements contenant différentes situations difficiles, telles que des encombrements d'arrière-plan, un fort bruit, des changements d'apparence, des occultations, des changements d'échelle, et l'apparition et la disparition de plusieurs objets en mouvement dans la scène. Dans une quatrième partie, nous avons présenté les expériences menées pour évaluer l'approche de reconnaissance de posture humaine proposée, qui est basée sur trois types de caractéristiques, à savoir les histogrammes de chaîne de codes, les moments de Krawtchouk, et les caractéristiques géométriques. Nous avons évalué les performances de ces caractéristiques dans un premier temps individuellement, puis en combinaison afin

de déterminer leur impact séparé et combiné sur la performance globale de la reconnaissance de posture. L'évaluation a été menée sur un ensemble de données contenant des silhouettes humaines prises à partir de différents angles de vue de la caméra. Enfin, dans une dernière partie, nous avons présenté un exemple d'application de notre système de vidéo surveillance intelligent proposé pour la détection d'un comportement humain anormal, qui est "la chute". Les résultats ont montré l'efficacité de notre système proposé à détecter ce genre d'événements très dangereux, et qui est très fréquent dans des conditions de mauvaise visibilité, telles que la nuit.

Conclusion générale et perspectives

Dans cette thèse, nous avons proposé un nouveau système pour la détection, le suivi et la reconnaissance automatique de posture de personnes en mouvement pour des applications de surveillance nocturne en extérieur. L'entrée de ce système est une séquence d'images prises par une caméra stationnaire infrarouge. Le système est donc capable de détecter la présence d'un humain dans la scène surveillée, puis d'alerter l'utilisateur lorsqu'un comportement anormal (par exemple, une chute) est observé.

Plus précisément, le système proposé comprend quatre phases successives qui sont la détection de personnes, le suivi, la reconnaissance de posture et l'analyse de comportement.

En effet, pour distinguer efficacement un être humain à partir d'autres objets non humains en mouvement, nous avons présenté deux approches différentes. La première est basée sur le calcul d'une fonction de similarité combinée qui utilise des informations de forme et d'apparence, et des informations spatiales et temporelles des objets en mouvement. Cette approche, malgré sa simplicité, a l'avantage de détecter des personnes en mouvement dans des séquences d'images infrarouges sans la nécessité d'un apprentissage préalable (a priori) d'un modèle mathématique (classifieur). Cependant, cette approche est applicable uniquement lorsque les personnes sont en mouvement dans les postures "Standing" (Debout) ou "Bending" (Penché). La deuxième approche présentée est basée sur la détection conjointe de la partie tête-épaule (ressemblant à la forme Omega Ω) et les deux jambes. Cette approche a été testée sur un ensemble de données contenant des humains et différentes espèces animales, et les résultats ont montré sa capacité à détecter des humains même en cas de présence de changements dans leur posture ou leur direction de mouvement.

Afin de suivre de manière robuste les personnes détectées, nous avons proposé une approche à base d'un filtre à particules et une combinaison adaptative de multiples

caractéristiques différentes, à savoir l'intensité, la texture, la vélocité de mouvement et la distance spatiale. Pour augmenter davantage la robustesse de cette méthode, nous avons également proposé une stratégie automatique pour la détection et le traitement des occlusions basée sur des règles heuristiques simples et l'histogramme de projection verticale en niveaux de gris. Cette approche a été testée sur des séquences d'images IR contenant différentes situations difficiles (telles que des encombrements d'arrière-plan, un fort bruit, des changements d'apparence, des occultations, etc.), et l'étude comparative avec certaines approches reportées dans la littérature a montré l'efficacité et la supériorité de notre approche proposée en termes de performance.

Pour identifier la posture des personnes suivies, nous avons proposé une nouvelle méthode basée sur un algorithme SVM multi-classe et une combinaison de trois caractéristiques différentes, qui sont : les moments de Krawtchouk, l'histogramme de chaîne de code et des caractéristiques géométriques. Ces caractéristiques ont été d'abord évaluées individuellement, puis en combinaison afin d'évaluer l'impact sur la performance globale de reconnaissance de posture. Les résultats obtenus sur un ensemble de données contenant des silhouettes humaines prises à partir de différents angles de vue de la caméra ont montré que, lorsqu'elles sont combinées, ces caractéristiques améliorent significativement le taux de reconnaissance de certaines classes de posture, telles que "Creeping" et "Crawling".

Dans la dernière phase, nous avons exploité les résultats de la reconnaissance de posture et les informations temporelles fournies par l'algorithme de suivi pour analyser le comportement des personnes détectées dans la scène surveillée. Cependant, afin de réduire l'effet des erreurs de segmentation sur la performance globale de l'analyse de comportement, nous avons utilisé un filtre à vote majoritaire pondéré par un noyau gaussien. La performance globale de l'analyse de comportement a été évaluée sur une séquence d'images infrarouges capturée dans un environnement nocturne extérieur réel. Les résultats expérimentaux obtenus ont montré la faisabilité et l'efficacité de notre méthode à analyser le comportement des personnes suivies et à détecter des événements anormaux tels les chutes, qui sont très fréquentes dans des conditions de très mauvaise visibilité.

Cependant, malgré les résultats prometteurs obtenus par le système proposé, certaines limites doivent être mentionnées. Tout d'abord, le système proposé est actuellement conçu pour surveiller uniquement plusieurs humains évoluant

indépendamment les uns des autres dans la scène surveillée, mais il n'est pas adéquat pour traiter certaines situations particulières, comme un groupe ou une foule de personnes. Une deuxième limitation est que, comme la plupart des systèmes basés sur la soustraction d'arrière-plan, la performance du système proposé est relativement sensible à la qualité des silhouettes extraites des objets en mouvement. Cette sensibilité peut être réduite à l'avenir en employant une technique de soustraction d'arrière-plan plus complexe, comme celles qui exploitent de multiples informations (Huerta et al., 2013; Noh and Jeon, 2013) au lieu de la seule information d'intensité utilisée dans ce travail. En outre, puisque le système proposé est conçu pour fonctionner dans des environnements extérieurs, nous avons l'intention, dans le cadre de nos recherches futures, de l'étendre aux environnements intérieurs, tels que des maisons, des lieux de travail, des chambres, etc. Dans ce cas, nous devons prendre en compte davantage de postures humaines que les cinq postures de base considérées dans ce travail.

Bibliographie

Abari, M.E., 2018. A Novel Pedestrian Detection Method Based on Combination of LBP, HOG, and Haar-Like Features, in: 2018 IEEE International Conference on Electro/Information Technology (EIT), pp. 0055–0066.

Adhikari, K., Bouchachia, H., Nait-Charif, H., 2017. Activity recognition for indoor fall detection using convolutional neural network, in: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), pp. 81–84.

Afsar, P., Cortez, P., Santos, H., 2017. Human Skeleton Detection from Semi-constrained Environment Video, in: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 384–389.

Amine Elforaici, M.E., Chaaraoui, I., Bouachir, W., Ouakrim, Y., Mezghani, N., 2018. Posture Recognition Using an RGB-D Camera: Exploring 3D Body Modeling and Deep Learning Approaches, in: 2018 IEEE Life Sciences Conference (LSC), pp. 69–72.

Andreone, L., Bellotti, F., De Gloria, A., Lauletta, R., 2005. SVM-based pedestrian recognition on near-infrared images, in: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, pp. 274–278.

Asha, C.S., Narasimhadhan, A.V., 2017. Robust infrared target tracking using discriminative and generative approaches. Infrared Physics & Technology 85, 114–127.

Asomaning, J., Haupt, S., Chae, M., Bressler, D.C., 2018. Recent developments in microwave-assisted thermal conversion of biomass for fuels and chemicals. Renewable and Sustainable Energy Reviews 92, 642–657.

Bai, X., Liu, W., Tu, Z., 2009. Integrating contour and skeleton for shape classification, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 360–367.

Bal, A., Alam, M.S., 2004. Dynamic target tracking with fringe-adjusted joint transform correlation and template matching. Applied Optics 43, 4874–4881.

Bali, S., Tyagi, S.S., 2018. A Review of Vision-Based Pedestrian Detection Techniques. International Journal of Advanced Studies of Scientific Research, 3(9).

Bao, C., Wu, Y., Ling, H., Ji, H., 2012. Real time robust L1 tracker using accelerated proximal gradient approach, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1830–1837.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110, 346–359.

Benabdeslem, K., Bennani, Y., 2006. Dendogram based SVM for multi-class classification, in: 28th International Conference on Information Technology Interfaces, 2006, pp. 173–178.

Benezeth, Y., Emile, B., Laurent, H., Rosenberger, C., 2010. Vision-Based System for Human Detection and Tracking in Indoor Environment. International Journal of Social Robotics 2, 41–52.

Benezeth, Y., Emile, B., Laurent, H., Rosenberger, C., 2008. A Real Time Human Detection System Based on Far Infrared Vision, in: Image and Signal Processing, pp. 76–84.

Bertozzi, M., Broggi, A., Gomez, C.H., Fedriga, R.I., Vezzoni, G., DelRose, M., 2007. Pedestrian Detection in Far Infrared Images based on the use of Probabilistic Templates, in: 2007 IEEE Intelligent Vehicles Symposium, pp. 327–332.

Bhusal, S., 2015. Object Detection and Tracking in Wide Area Surveillance Using Thermal Imagery. UNLV Theses, Dissertations, Professional Papers, and Capstones. 2517.
Binelli, E., Broggi, A., Fascioli, A., Ghidoni, S., Grisleri, P., Graf, T., Meinecke, M., 2005. A modular tracking system for far infrared pedestrian recognition, in: IEEE Proceedings. Intelligent Vehicles Symposium, 2005, pp. 759–764.
Bishop, C., 2006. Pattern Recognition and Machine Learning, Information Science and Statistics. Springer-Verlag, New York.
Boulay, B., Brémond, F., Thonnat, M., 2006. Applying 3D human model in a posture recognition system. Pattern Recognition Letters, Vision for Crime Detection and Prevention 27, 1788–1796.
Bourennane, S., Fossati, C., 2012. Comparison of shape descriptors for hand posture recognition in video. Signal, Image and Video Processing 6, 147–157.
Brehar, R., Vancea, C., Nedevschi, S., 2014. Pedestrian detection in infrared images using Aggregated Channel Features, in: 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 127–132.
Brulin, D., Benezeth, Y., Courtial, E., 2012. Posture Recognition Based on Fuzzy Logic for Home Monitoring of the Elderly. IEEE Transactions on Information Technology in Biomedicine 16, 974–982.
Buccolieri, F., Distante, C., Leone, A., 2005. Human posture recognition using active contours and radial basis function neural network, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005, pp. 213–218.
Budzan, S., Wyżgolik, R., 2015. Remarks on noise removal in infrared images. Measurement Automation Monitoring 61.
Chen, H.-S., Chen, H.-T., Chen, Y.-W., Lee, S.-Y., 2006. Human action recognition using star skeleton, in: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, pp. 171–178.
Chen, J., Lin, Y., Huang, D., Zhang, J., 2020. Robust tracking algorithm for infrared target via correlation filter and particle filter. Infrared Physics & Technology 111, 103516.
Chen, S., Akselrod, P., Zhao, B., Perez Carrasco, J.A., Linares-Barranco, B., Culurciello, E., 2012. Efficient Feedforward Categorization of Objects and Human Postures with Address-Event Image Sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 302–314.
Chen, Y., Liu, X., Huang, Q., 2008. Real-time detection of rapid moving infrared target on variation background. Infrared Physics & Technology 51, 146–151.
Choi, J.-W., Moon, D., Yoo, J.-H., 2015. Robust Multi-person Tracking for Real-Time Intelligent Video Surveillance. ETRI Journal 37, 551–561.
Choi, W., Pantofaru, C., Savarese, S., 2011. Detecting and tracking people using an RGB-D camera via multiple detector fusion, in: 2011 IEEE International Conference on Computer Vision Workshops, pp. 1076–1083.
Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L., 2000. A System for Video Surveillance and Monitoring. VSAM final report 2000, 69.
Conaire, C.O., O'Connor, N.E., Cooke, E., Smeaton, A.F., 2006. Comparison of Fusion Methods for Thermo-Visual Surveillance Tracking, in: 2006 9th International Conference on Information Fusion, pp. 1–7.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297.

Cruz, F.C., Nogueira, G.T., Costa, L.F.L., Frateschi, N.C., Viscovini, R.C., Pereira, D., 2007. Continuous and Pulsed THz generation with molecular gas lasers and photoconductive antennas gated by femtosecond pulses, in: 2007 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference, pp. 446–449.

Cucchiara, R., Grana, C., Prati, A., Vezzani, R., 2005. Probabilistic posture classification for Human-behavior analysis. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 35, 42–54.

Dai, C., Zheng, Y., Li, X., 2007. Pedestrian detection and tracking in infrared imagery using shape and appearance. Computer Vision and Image Understanding, Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum 106, 288–299.

Dai, X., Duan, Y., Hu, J., Liu, S., Hu, C., He, Y., Chen, D., Luo, C., Meng, J., 2019. Near infrared nighttime road pedestrians recognition based on convolutional neural network. Infrared Physics & Technology 97, 25–32.

Dalal, N., Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893.

D'Amico, A., Natale, C.D., Castro, F.L., Iarossi, S., Catini, A., Martinelli, E., 2009. Volatile Compounds Detection by IR Acousto-Optic Detectors, in: Unexploded Ordnance Detection and Mitigation. Springer Netherlands, Dordrecht, pp. 21–59.

Dash, M.C., Dash, S.P., 2009. Fundamentals of ecology. Tata McGraw-Hill Education.

Davis, J.W., Keck, M.A., 2005. A Two-Stage Template Approach to Person Detection in Thermal Imagery, in: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1, pp. 364–369.

Davis, J.W., Sharma, V., 2007. Background-subtraction using contour-based fusion of thermal and visible imagery. Computer Vision and Image Understanding, Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum 106, 162–182.

Davis, J.W., Sharma, V., Tyagi, A., Keck, M., 2009. Human Detection and Tracking, in: Li, S.Z., Jain, A. (Eds.), Encyclopedia of Biometrics. Springer US, Boston, MA, pp. 708–712.

Ding, M., Chen, W. H., Cao, Y. F., 2022. Thermal infrared single-pedestrian tracking for advanced driver assistance system. IEEE Transactions on Intelligent Vehicles 8, 814-824.

Ding, W., Hu, B., Liu, H., Wang, X., Huang, X., 2020. Human posture recognition based on multiple features and rule learning. International Journal of Machine Learning and Cybernetics 11, 2529–2540.

Diraco, G., Leone, A., Siciliano, P., 2013. Human posture recognition with a time-of-flight 3D sensor for in-home applications. Expert Systems with Applications 40, 744–751.

Dong, J., Ge, J., Luo, Y., 2007. Nighttime Pedestrian Detection with Near Infrared using Cascaded Classifiers, in: 2007 IEEE International Conference on Image Processing, pp. VI-185-VI–188.

Dong Xia, Hao Sun, Zhenkang Shen, 2010. Real-time infrared pedestrian detection based on multi-block LBP, in: 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), pp. V12-139-V12-142.

Feng, G., Lin, Q., 2010. Design of elder alarm system based on body posture reorganization, in: Security and Identification 2010 International Conference on Anti-Counterfeiting, pp. 249–252.
Fletcher, T., 2009. Support Vector Machines Explained. Tutorial paper 1–19.
Flir Systems, 2021. The Ultimate Infrared Handbook for R&D Professionals - A Resource Guide for Using Infrared in the Research and Development Industry [WWW Document]. URL https://www.flirmedia.com/MMC/THG/Brochures/T559243/T559243_EN.pdf (accessed 7.27.22).
Flir Systems, 2018. Cooled versus uncooled thermal cameras for long-range surveillance [WWW Document]. URL http://www.flirmedia.com/MMC/CVS/Tech_Notes/TN_0005_EN.pdf (accessed 9.21.22).
Gade, R., Moeslund, T.B., 2014. Thermal cameras and applications: a survey. Machine Vision and Applications 25, 245–262.
Gao, J., Ling, H., Hu, W., Xing, J., 2014. Transfer Learning Based Visual Tracking with Gaussian Processes Regression, in: Computer Vision – ECCV 2014, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 188–203.
Gao, S.J., Jhang, S.T., 2016. Infrared Target Tracking Using Multi-Feature Joint Sparse Representation, in: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, RACS '16. Association for Computing Machinery, New York, NY, USA, pp. 40–45.
García-Martín, Á., Martínez, J.M., 2012. On collaborative people detection and tracking in complex scenarios. Image and Vision Computing 30, 345–354.
Gautam, G., Choudhary, K., Chatterjee, S., Kolekar, M.H., 2017. Facial expression recognition using krawtchouk moments and support vector machine classifier, in: 2017 Fourth International Conference on Image Information Processing (ICIIP), pp. 1–6.
Ge, J., Luo, Y., Tei, G., 2009. Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems. IEEE Transactions on Intelligent Transportation Systems 10, 283–298.
Girondel, V., Bonnaud, L., Caplier, A., Rombaut, M., 2005. Belief theory-based classifiers comparison for static human body postures recognition in video. International Journal of Signal Processing 2, 29.
Goldmann, L., Karaman, M., Sikora, T., 2004. Human body posture recognition using MPEG-7 descriptors, in: Visual Communications and Image Processing 2004, International Society for Optics and Photonics, pp. 177–188.
Govardhan, P., Pati, U.C., 2014. NIR image based pedestrian detection in night vision with cascade classification and validation, in: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp. 1435–1438.
Goyal, K., Singhai, J., 2018. Review of background subtraction methods using Gaussian mixture model for video surveillance systems. Artificial Intelligence Review 50, 241–259.
Grenn, M.W., Vizgaitis, J., Pellegrino, J.G., Perconti, P., 2012. Infrared Camera and Optics for Medical Applications, in: Medical Infrared Imaging: Principles and Practices. Edited by N. A. Diakides and J. D. Bronzino, CRC Press, Taylor and Francis Group.

Guo, L., Ge, P.-S., Zhang, M.-H., Li, L.-H., Zhao, Y.-B., 2012. Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine. Expert Systems with Applications 39, 4274–4286.
Haken, H., Wolf, H.C., 2013. Molecular Physics and Elements of Quantum Chemistry: Introduction to Experiments and Theory. Springer Science & Business Media.
Hamamatsu Photonics, 2021. NIR and SWIR Questions and Answers [WWW Document]. URL https://hub.hamamatsu.com/us/en/ask-engineer/nir-and-swir-questions-and-answers/index.html (accessed 6.5.22).
Han, T.Y., Song, B.C., 2016. Night vision pedestrian detection based on adaptive preprocessing using near infrared camera, in: 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1–3.
Haq, E.U., Jianjun, H., Li, K., Haq, H.U., 2020. Human detection and tracking with deep convolutional neural networks under the constrained of noise and occluded scenes. Multimedia Tools and Applications 79, 30685–30708.
Haritaoglu, I., Harwood, D., Davis, L.S., 2000. W4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 809–830.
Haritaoglu, I., Harwood, D., Davis, L.S., 1998. Ghost: a human body part labeling system using silhouettes, in: Proceedings. Fourteenth International Conference on Pattern Recognition, pp. 77–82 vol.1.
He, Y.-J., Li, M., Zhang, J., Yao, J.-P., 2015. Infrared target tracking via weighted correlation filter. Infrared Physics & Technology 73, 103–114.
Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2012. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels, in: Computer Vision – ECCV 2012, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 702–715.
Heo, D., Lee, E., Ko, B.C., 2018. Pedestrian Detection at Night Using Deep Neural Networks and Saliency Maps. Electronic Imaging 2018, 060403-1-060403–9.
Herrmann, C., Müller, T., Willersinn, D., Beyerer, J., 2016. Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs, in: Electro-Optical and Infrared Systems: Technology and Applications XIII, pp. 166-173.
Martin, G., 2015. High Performance SWIR Imaging Cameras. Raptor Photonics White Papers; Raptor Photonics Ltd.: Milbrook, Larne, UK.
Hitchcock, R.T., 2004. Radio-frequency and Microwave Radiation. American Industrial Hygiene Association.
Hoa, N.T., Bui, T.D., 2016. Classifying human body postures by a two-neuron fuzzy neural network, in: 2016 IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp. 142–146.
Hu, J.-S., Su, T.-M., Lin, P.-C., 2007. 3-D Human Posture Recognition System Using 2-D Shape Features, in: Proceedings 2007 IEEE International Conference on Robotics and Automation, pp. 3933–3938.
Huerta, I., Amato, A., Roca, X., Gonzàlez, J., 2013. Exploiting multiple cues in motion segmentation based on background subtraction. Neurocomputing, Special issue: Behaviours in video 100, 183–196.
Idris, M.I., Zabidi, A., Yassin, I.M., Ali, M.S.A.M., 2015. Human posture recognition using android smartphone and artificial neural network, in: 2015 IEEE 6th Control and System Graduate Research Colloquium (ICSGRC), pp. 120–124.

Isard, M., Blake, A., 1998. CONDENSATION—Conditional Density Propagation for Visual Tracking. International Journal of Computer Vision 29, 5–28.
Jeon, E.S., Kim, J.H., Hong, H.G., Batchuluun, G., Park, K.R., 2016. Human Detection Based on the Generation of a Background Image and Fuzzy System by Using a Thermal Camera. Sensors 16, 453.
Jia, X., Lu, H., Yang, M.-H., 2012. Visual tracking via adaptive structural local sparse appearance model, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1822–1829.
John, V., Mita, S., Liu, Z., Qi, B., 2015. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks, in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA), pp. 246–249.
Jones, B.F., 1998. A reappraisal of the use of infrared thermal image analysis in medicine. IEEE Transactions on Medical Imaging 17, 1019–1027.
Joshi, R.C., Joshi, M., Singh, A.G., Mathur, S., 2018. Object Detection, Classification and Tracking Methods for Video Surveillance: A Review, in: 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–7.
Juang, C.-F., Chang, C.-M., 2007. Human Body Posture Classification by a Neural Fuzzy Network and Home Care System Application. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 37, 984–994.
Juang, C.-F., Chang, C.-M., Wu, J.-R., Lee, D., 2009. Computer Vision-Based Human Body Segmentation and Posture Estimation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 39, 119–133.
Juang, C.-F., Chen, T.-C., Du, W.-C., 2014. Human Body 3D Posture Estimation Using Significant Points and Two Cameras. The Scientific World Journal 2014, 17.
Jüngling, K., Arens, M., 2010. Pedestrian tracking in infrared from moving vehicles, in: 2010 IEEE Intelligent Vehicles Symposium, pp. 470–477.
Kancharla, T., Kharade, P., Gindi, S., Kutty, K., Vaidya, V.G., 2011. Edge based segmentation for pedestrian detection using NIR camera, in: 2011 International Conference on Image Information Processing, pp. 1–6.
Kang, H.-G., Lee, S.-H., 2016. Human body posture recognition with discrete cosine transform, in: 2016 International Conference on Big Data and Smart Computing (BigComp), pp. 423–426.
Kim, D., Jansen, R.A., Windhorst, R.A., 2017. Analysis of the Intrinsic Mid-infrared L band to Visible–Near-infrared Flux Ratios in Spectral Synthesis Models of Composite Stellar Populations. The Astrophysical Journal 840, 28.
Kim, D.-E., Kwon, D.-S., 2015. Pedestrian detection and tracking in thermal images using shape features, in: 2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 22–25.
Kim, T., Kim, S., 2018. Pedestrian detection at night time in FIR domain: Comprehensive study about temperature and brightness and new benchmark. Pattern Recognition 79, 44–54.
Kim, Y., Bang, H., 2018. Introduction to Kalman Filter and Its Applications. IntechOpen.
Kumar, K.S.C., 2013. Phase-edge based approach for pedestrian segmentation using NIR camera and tracking for driver assistance, in: 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), pp. 214–218.

Kuptametee, C., Aunsri, N., 2022. A review of resampling techniques in particle filtering framework. Measurement, 110836

Kwan, C., Chou, B., Yang, J., Tran, T., 2019. Deep Learning Based Target Tracking and Classification for Infrared Videos Using Compressive Measurements. Journal of Signal and Information Processing 10, 167.

Lahouli, I., Karakasis, E., Haelterman, R., Chtourou, Z., Cubber, G.D., Gasteratos, A., Attia, R., 2018. Hot spot method for pedestrian detection using saliency maps, discrete Chebyshev moments and support vector machine. IET Image Processing 12, 1284–1291.

Larsson, F., Felsberg, M., 2011. Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition, in: Image Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 238–249.

Le, T.-L., Nguyen, M.-Q., Nguyen, T.-T.-M., 2013. Human posture recognition using human skeleton provided by Kinect, in: 2013 International Conference on Computing, Management and Telecommunications (ComManTel), pp. 340–345.

Lee, J.H., Choi, J.-S., Jeon, E.S., Kim, Y.G., Le, T.T., Shin, K.Y., Lee, H.C., Park, K.R., 2015. Robust Pedestrian Detection by Combining Visible and Thermal Infrared Cameras. Sensors 15, 10580–10615.

Lee, S.-H., Kim, J.-H., Choi, K.P., Sim, J.-Y., Kim, C.-S., 2014. Video saliency detection based on spatiotemporal feature learning, in: 2014 IEEE International Conference on Image Processing (ICIP), pp. 1120–1124.

Lee, S.J., Shah, G., Bhattacharya, A.A., Motai, Y., 2012. Human tracking with an infrared camera using a curve matching framework. EURASIP Journal on Advances in Signal Processing 2012, 99.

Lee, Y., Chan, Y., Fu, L., Hsiao, P., 2015. Near-Infrared-Based Nighttime Pedestrian Detection Using Grouped Part Models. IEEE Transactions on Intelligent Transportation Systems 16, 1929–1940.

Li, B., Han, C., Bai, B., 2019. Hybrid approach for human posture recognition using anthropometry and BP neural network based on Kinect V2. EURASIP Journal on Image and Video Processing 2019, 8.

Li, C.-C., Chen, Y.-Y., 2006. Human Posture Recognition by Simple Rules, in: 2006 IEEE International Conference on Systems, Man and Cybernetics, pp. 3237–3240.

Li, H., Sun, Q., 2013. The recognition of moving human body posture based on combined neural network, in: IEEE Conference Anthology, pp. 1–5.

Li, J., Gong, W., 2010. Real Time Pedestrian Tracking using Thermal Infrared Imagery. JCP 5, 1606–1613.

Li, J., Gong, W., Li, W., Liu, X., 2010. Robust pedestrian detection in thermal infrared imagery using the wavelet transform. Infrared Physics & Technology 53, 267–273.

Li, J., Wang, Y., 2009. Pedestrian tracking in infrared image sequences using wavelet entropy features, in: 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA), pp. 288–291.

Li, J., Zhang, F., Wei, L., Yang, T., Lu, Z., 2017. Nighttime Foreground Pedestrian Detection Based on Three-Dimensional Voxel Surface Model. Sensors 17, 2354.

Li Zhang, Wu, B., Nevatia, R., 2007. Pedestrian Detection in Infrared Images based on Local Shape Features, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Lim, C.C., Basah, S.N., Ali, M.A., Fook, C.Y., 2018. Wearable Posture Identification System for Good Sitting Position. Journal of Telecommunication, Electronic and Computer Engineering 10, 135–140.
Lin, Y., Chan, Y., Chuang, L., Fu, L., Huang, S., Hsiao, P., Luo, M., 2011. Near-infrared based nighttime pedestrian detection by combining multiple features, in: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1549–1554.
Liu, H., Luo, J., Wu, P., Xie, S., Li, H., 2016. People detection and tracking using RGB-D cameras for mobile robots. International Journal of Advanced Robotic Systems 13, 172988141665774.
Liu, J., Liu, Y., Cui, Y., Chen, Y.Q., 2013. Real-time human detection and tracking in complex environments using single RGBD camera, in: 2013 IEEE International Conference on Image Processing, pp. 3088–3092.
Liu, Q., Lu, X., He, Z., Zhang, C., Chen, W.-S., 2017. Deep convolutional neural networks for thermal infrared object tracking. Knowledge-Based Systems 134, 189–198.
Liu, Q., Zhuang, J., Ma, J., 2013. Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems. Infrared Physics & Technology 60, 288–299.
Liu, R., Yang, M., 2012. Tracking Multiple Feature in Infrared Image with Mean-Shift, in: Advanced Intelligent Computing, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 194–201.
Liu, Y., 2021. Human motion image detection and tracking method based on Gaussian mixture model and CAMSHIFT. Microprocessors and Microsystems 82, 103843.
Liu, Y., Huang, S., Lu, C., Chang, F., Lin, P., 2017. Thermal pedestrian detection using block LBP with multi-level classifier, in: 2017 International Conference on Applied System Innovation (ICASI), pp. 602–605.
Luna, C.A., Losada-Gutierrez, C., Fuentes-Jimenez, D., Fernandez-Rincon, A., Mazo, M., Macias-Guarasa, J., 2017. Robust people detection using depth information from an overhead Time-of-Flight camera. Expert Systems with Applications 71, 240–256.
Ma, N., Wu, Z., Cheung, Y. M., Guo, Y., Gao, Y., Li, J., Jiang, B., 2022. A Survey of Human Action Recognition and Posture Prediction. Tsinghua Science and Technology 27, 973–1001.
Mehta, R., Egiazarian, K., 2016. Dominant Rotated Local Binary Patterns (DRLBP) for texture classification. Pattern Recognition Letters 71, 16–22.
Mesbah, A., El Mallahi, M., Lakhili, Z., Qjidaa, H., Berrahou, A., 2016. Fast and accurate algorithm for 3D local object reconstruction using Krawtchouk moments, in: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS), pp. 1–6.
Miezianko, R., Pokrajac, D., 2008. People detection in low resolution infrared videos, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6.
Mikolajczyk, K., Schmid, C., Zisserman, A., 2004. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors, in: Computer Vision - ECCV 2004, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 69–82.
Miron, A., Besbes, B., Rogozan, A., Ainouz, S., Bensrhair, A., 2012. Intensity self similarity features for pedestrian detection in Far-Infrared images, in: 2012 IEEE Intelligent Vehicles Symposium, pp. 1120–1125.

Munaro, M., Lewis, C., Chambers, D., Hvass, P., Menegatti, E., 2016. RGB-D Human Detection and Tracking for Industrial Environments, in: Intelligent Autonomous Systems 13, Springer, pp. 1655–1668.
Nadeem, A., Jalal, A., Kim, K., 2021. Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model. Multimedia Tools and Applications 80, 21465–21498.
Nanda, H., Davis, L., 2002. Probabilistic template based pedestrian detection in infrared videos, in: IEEE Intelligent Vehicle Symposium, 2002, pp. 15–20 vol.1.
Negied, N.K., Hemayed, E.E., Fayek, M.B., 2015. Pedestrians' detection in thermal bands – Critical survey. Journal of Electrical Systems and Information Technology 2, 141–148.
Nguyen, D.T., Li, W., Ogunbona, P., 2010. Human detection using local shape and Non-Redundant binary patterns, in: 11th International Conference on Control Automation Robotics Vision, pp. 1145–1150.
Nguyen, D.T., Li, W., Ogunbona, P.O., 2016. Human detection from images and videos: A survey. Pattern Recognition 51, 148–175.
Noh, S., Jeon, M., 2013. A New Framework for Background Subtraction Using Multiple Cues, in: 2012 Asian Conference on Computer Vision, Lecture Notes in Computer Science. Springer, pp. 493–506.
Nummiaro, K., Koller-Meier, E., Van Gool, L., 2003. An adaptive color-based particle filter. Image and Vision Computing 21, 99–110.
Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29, 51–59.
Olmeda, D., Escalera, A. de la, Armingol, J.M., 2012. Contrast invariant features for human detection in far infrared images, in: 2012 IEEE Intelligent Vehicles Symposium, pp. 117–122.
Olmeda, D., Escalera, A. de la, Armingol, J.M., 2011. Far infrared pedestrian detection and tracking for night driving. Robotica 29, 495–505.
O'Malley, R., Jones, E., Glavin, M., 2010. Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation. Infrared Physics & Technology 53, 439–449.
Opgal, 2021. Intro to IR (Part 2): Cooled vs. uncooled cameras, sensitivity, resolution, frame rate [WWW Document]. URL https://www.opgal.com/blog/thermal-cameras/intro-to-ir-part-2-cooled-vs-uncooled-cameras-sensitivity-resolution-frame-rate/ (accessed 9.21.22).
Oshima, N., Saitoh, T., Konishi, R., 2006. Real Time Mean Shift Tracking using Optical Flow Distribution, in: 2006 SICE-ICASE International Joint Conference, pp. 4316–4320.
Pang, Y., Yuan, Y., Li, X., Pan, J., 2011. Efficient HOG human detection. Signal Processing 91, 773–781.
Paravati, G., Esposito, S., 2014. Relevance-Based Template Matching for Tracking Targets in FLIR Imagery. Sensors (Basel) 14, 14106–14130.
Park, J., Chen, J., Cho, Y.K., Kang, D.Y., Son, B.J., 2020. CNN-Based Person Detection Using Infrared Images for Night-Time Intrusion Warning Systems. Sensors 20, 34.
Patel, A., 2013. Chapter 40 - Anesthesia for Laser Airway Surgery, in: Hagberg, C.A. (Ed.), Benumof and Hagberg's Airway Management (Third Edition). W.B. Saunders, Philadelphia, pp. 824-858.e4.

Pellegrini, S., Iocchi, L., 2008. Human Posture Tracking and Classification through Stereo Vision and 3D Model Matching. EURASIP Journal on Image and Video Processing 2008, 1–12.
Pérez, P., Hue, C., Vermaak, J., Gangnet, M., 2002. Color-Based Probabilistic Tracking, in: Computer Vision — ECCV 2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 661–675.
Picart, P., 2015. New Techniques in Digital Holography. John Wiley & Sons.
Piniarski, K., Pawłowski, P., 2016. Multi-branch classifiers for pedestrian detection from infrared night and day images, in: 2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 248–253.
Portmann, J., Lynen, S., Chli, M., Siegwart, R., 2014. People detection and tracking from aerial thermal views, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 1794–1800.
Pradhan, R., Kumar, S., Agarwal, R., Pradhan, M.P., Ghose, M.K., 2010. Contour Line Tracing Algorithm for Digital Topographic Maps. International Journal of Image Processing 4, 156–163.
Qi, B., John, V., Liu, Z., Mita, S., 2016. Pedestrian detection from thermal images: A sparse representation based approach. Infrared Physics & Technology 76, 157–167.
Ran, Y., Weiss, I., Zheng, Q., Davis, L.S., 2007. Pedestrian Detection via Periodic Motion Analysis. International Journal of Computer Vision 71, 143–160.
Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517–6525.
Rocha, L., Velho, L., Carvalho, P.C.P., 2004. Motion reconstruction using moments analysis, in: Proceedings. 17th Brazilian Symposium on Computer Graphics and Image Processing, pp. 354–361.
Rowan-Robinson, M., 2013. Night Vision: Exploring the Infrared Universe. Cambridge University Press.
Sahani, S.K., Adhikari, G., Das, B., 2011. A fast template matching algorithm for aerial object tracking, in: 2011 International Conference on Image Information Processing, pp. 1–6.
Satake, J., Miura, J., 2009. Robust Stereo-Based Person Detection and Tracking for a Person Following Robot, in: ICRA Workshop on People Detection and Tracking, pp. 1–10.
Satpathy, A., Jiang, X., Eng, H.-L., 2013. Human detection using Discriminative and Robust Local Binary Pattern, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2376–2380.
Seemann, E., Leibe, B., Mikolajczyk, K., Schiele, B., 2005. An Evaluation of Local Shape-Based Features for Pedestrian Detection, in: Procedings of the British Machine Vision Conference, pp. 5.1-5.10.
Shahzad, A. R., Jalal, A., 2021. A smart surveillance system for pedestrian tracking and counting using template matching, in: 2021 International Conference on Robotics and Automation in Industry (ICRAI), pp. 1-6.
Shalaby, M., Vicario, C., Hauri, C.P., 2015. High-performing nonlinear visualization of terahertz radiation on a silicon charge-coupled device. Nature Communications 6, 8439.
Shechtman, E., Irani, M., 2007. Matching Local Self-Similarities across Images and Videos, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images, in: 2011 IEEE Computer Vision and Pattern Recognition, pp. 1297–1304.

Shuang, Z., Yu-Ping, Q., 2012. Mean-Shift Algorithm Apply for Infrared Imaging Tracking. AASRI Procedia, AASRI Conference on Computational Intelligence and Bioinformatics 1, 52–57.

Singh, M., Basu, A., Mandal, M.K., 2008. Human Activity Recognition Based on Silhouette Directionality. IEEE Transactions on Circuits and Systems for Video Technology 18, 1280–1292.

Smith, S.M., Brady, J.M., 1997. SUSAN—A New Approach to Low Level Image Processing. International Journal of Computer Vision 23, 45–78.

Soga, M., Hiratsuka, S., Fukamachi, H., Ninomiya, Y., 2008. Pedestrian Detection for a Near Infrared Imaging System, in: 2008 11th International IEEE Conference on Intelligent Transportation Systems, pp. 1167–1172.

Soundrapandiyan, R., Chandra Mouli, P.V.S.S.R., 2018. An Approach to Adaptive Pedestrian Detection and Classification in Infrared Images Based on Human Visual Mechanism and Support Vector Machine. Arabian Journal for Science and Engineering 43, 3951–3963.

Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A., 2006. Pedestrian Detection using Infrared images and Histograms of Oriented Gradients, in: 2006 IEEE Intelligent Vehicles Symposium, pp. 206–212.

Sun, H., Wang, C., Wang, B., El-Sheimy, N., 2011. Pyramid binary pattern features for real-time pedestrian detection from infrared videos. Neurocomputing 74, 797–804.

Tahir, N.M., Hussain, A., Samad, S.A., Husain, H., Jin, A.T.B., 2007. On the use of advanced correlation filters for human posture recognition. Journal of Applied Sciences 7, 2947–2956.

Taillet, R., Villain, L., Febvre, P., 2018. Dictionnaire de physique, 4e ed. De Boeck Supérieur.

Takahashi, K., Naemura, M., 2005. Remarks on human body posture estimation using neural network and Kalman filter, in: 2005 IEEE International Conference on Systems, Man and Cybernetics, pp. 2495–2500.

Tan, P., Huang, J., Liu, K., Xiong, Y., Fan, M., 2012. Terahertz radiation sources based on free electron lasers and their applications. Science China Information Sciences 55, 1–15.

Tanaka, S., Motoi, K., Nogawa, M., Yamakoshi, K., 2004. A new portable device for ambulatory monitoring of human posture and walking velocity using miniature accelerometers and gyroscope, in: The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2283–2286.

Teledyne DALSA, 2021. CCD vs CMOS. URL https://www.teledynedalsa.com/en/learn/knowledge-center/ccd-vs-cmos/ (accessed 5.22.22).

Teutsch, M., Mueller, T., Huber, M., Beyerer, J., 2014. Low Resolution Person Detection with a Moving Thermal Infrared Camera by Hot Spot Classification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 209–216.

Thieuleux, C., Boualleg, M., Candy, J.-P., Veyre, L., Basset, J.-M., 2011. Method for preparing a structured porous material comprising nanoparticles of metal 0 imbedded in the walls thereof. U.S. Patent Application No. 13/122,420.

Ukida, H., Kaji, S., Tanimoto, Y., Yamamoto, H., 2006. Human Motion Capture System using Color Markers and Silhouette, in: 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, pp. 151–156.
Viola, P., Jones, M.J., Snow, D., 2005. Detecting Pedestrians Using Patterns of Motion and Appearance. International Journal of Computer Vision 63, 153–161.
Walk, S., Majer, N., Schindler, K., Schiele, B., 2010. New features and insights for pedestrian detection, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1030–1037.
Wang, C., Zhou, J., Kou, P., Luo, Z., Zhang, Y., 2012. Identification of shaft orbit for hydraulic generator unit using chain code and probability neural network. Applied Soft Computing 12, 423–429.
Wang, J., Chen, D., Chen, H., Yang, J., 2012. On pedestrian detection and tracking in infrared videos. Pattern Recognition Letters 33, 775–785.
Wang, J., Huang, Z., Zhang, W., Patil, A., Patil, K., Zhu, T., Shiroma, E.J., Schepps, M.A., Harris, T.B., 2016. Wearable sensor based human posture recognition, in: 2016 IEEE International Conference on Big Data (Big Data), pp. 3432–3438.
Wang, L., Hu, W., Tan, T., 2003. Recent developments in human motion analysis. Pattern Recognition 36, 585–601.
Wang, W.-J., Chang, J.-W., Haung, S.-F., Wang, R.-J., 2016. Human Posture Recognition Based on Images Captured by the Kinect Sensor. International Journal of Advanced Robotic Systems 13, 54.
Wang, X., Liu, L., Tang, Z., 2009. Infrared human tracking with improved mean shift algorithm based on multicue fusion. Applied Optics 48, 4201–4212.
Wang, X., Tang, Z., 2010. Modified particle filter-based infrared pedestrian tracking. Infrared Physics & Technology 53, 280–287.
Wang, X., Xu, L., Ning, C., 2019. Multi-feature local sparse representation for infrared pedestrian tracking. KSII Transactions on Internet and Information Systems (TIIS) 13, 1464–1480.
Wientapper, F., Ahrens, K., Wuest, H., Bockholt, U., 2009. Linear-projection-based classification of human postures in time-of-flight data, in: 2009 IEEE International Conference on Systems, Man and Cybernetics, pp. 559–564.
Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P., 1997. Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 780–785.
Wu, B., Nevatia, R., 2007. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. International Journal of Computer Vision 75, 247–266.
Xie, F., Xu, G., Cheng, Y., Tian, Y., 2011. Human body and posture recognition system based on an improved thinning algorithm. IET Image Processing 5, 420–428.
Xu, F., Liu, X., Fujimura, K., 2005. Pedestrian detection and tracking with night vision. IEEE Transactions on Intelligent Transportation Systems 6, 63–71.
Xu, Y., Chen, J., Yang, Q., Guo, Q., 2019. Human Posture Recognition and fall detection Using Kinect V2 Camera, in: 2019 Chinese Control Conference (CCC), pp. 8488-8493.
Xu, Y., Wan, M., Chen, Q., Qian, W., Ren, K., Gu, G., 2021. Hierarchical convolution fusion-based adaptive Siamese network for infrared target tracking. IEEE Transactions on Instrumentation and Measurement 70, 1-12.

Yadong Mu, Shuicheng Yan, Yi Liu, Huang, T., Bingfeng Zhou, 2008. Discriminative local binary patterns for human detection in personal album, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Yajun Fang, Yamada, K., Ninomiya, Y., Horn, B.K.P., Masaki, I., 2004. A shape-independent method for pedestrian detection with far-infrared images. IEEE Transactions on Vehicular Technology 53, 1679–1697.

Yang, J., Marler, T., Rahmatallah, S., 2011. Multi-objective optimization-based method for kinematic posture prediction: development and validation. Robotica 29, 245.

Yang, M., Kpalma, K., Ronsin, J., 2008. A Survey of Shape Feature Extraction Techniques, in: Pattern Recognition. IN-TECH, pp. 43–90.

Yang, T., Fu, D., Pan, S., 2017. Pedestrian tracking for infrared image sequence based on trajectory manifold of spatio-temporal slice. Multimedia Tools and Applications 76, 11021–11035.

Yao, S., Pan, S., Wang, T., Zheng, C., Shen, W., Chong, Y., 2015a. A new pedestrian detection method based on combined HOG and LSS features. Neurocomputing 151, 1006–1014.

Yao, S., Wang, T., Shen, W., Pan, S., Chong, Y., Ding, F., 2015b. Feature Selection and Pedestrian Detection Based on Sparse Representation. PLOS ONE 10, e0134242.

Yasuno, M., Yasuda, N., Aoki, M., 2004. Pedestrian Detection and Tracking in Far Infrared Images, in: 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 125–125.

Younsi, M., Diaf, M., Siarry, P., 2020. Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera. Expert Systems with Applications 146, 113171.

Younsi, M., Diaf, M., Siarry, P., 2023a. Comparative study of orthogonal moments for human postures recognition. Engineering Applications of Artificial Intelligence 120, 105855.

Younsi, M., Yesli, S., Diaf, M., 2023b. Depth-based human action recognition using histogram of templates. Multimedia Tools and Applications, 1-35.

Yu, M., Rhuma, A., Naqvi, S.M., Wang, L., Chambers, J., 2012. A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment. IEEE Transactions on Information Technology in Biomedicine 16, 1274–1286.

Yu, Miao, Rhuma, A., Naqvi, S.M., Wang, L., Chambers, J., 2012. A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment. IEEE Transactions on Information Technology in Biomedicine 16, 1274–1286.

Yu, T., Mo, B., Liu, F., Qi, H., Liu, Y., 2019. Robust thermal infrared object tracking with continuous correlation filters and adaptive feature fusion. Infrared Physics & Technology 98, 69–81.

Yuan, D., Shu, X., Liu, Q., Zhang, X., He, Z., 2023. Robust thermal infrared tracking via an adaptively multi-feature fusion model. Neural Computing and Applications 35, 3423-3434.

Yun, S., Kim, S., 2019. Robust infrared target tracking using thermal information in mean-shift, in: Pattern Recognition and Tracking XXX, p. 1099509.

Žalik, B., Mongus, D., Lukač, N., 2015. A universal chain code compression method. Journal of Visual Communication and Image Representation 29, 8–15.

Zerrouki, N., Houacine, A., 2018. Combined curvelets and hidden Markov models for human fall detection. Multimedia Tools and Applications 77, 6405–6424.
Zerrouki, N., Houacine, A., 2014. Automatic Classification of Human Body Postures Based on the Truncated SVD. Journal of Advances in Computer Networks 2, 58–62.
Zhang, T., Liu, S., Xu, C., Liu, B., Yang, M. H., 2017. Correlation particle filter for visual tracking. IEEE Transactions on Image Processing 27, 2676–2687.
Zhao, X., He, Z., Zhang, S., Liang, D., 2015. Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification. Pattern Recognition 48, 1947–1960.
Zin, T.T., Tin, P., Hama, H., 2011. Pedestrian detection based on hybrid features using near infrared images. International Journal of Innovative Computing, Information and Control 7, 5015–5025.