Détection et suivi de personnes par vision omnidirectionnelle: approches 2D et 3D

Marouane Boui

université PARIS-SACLAY

: 2017SACLE035

Thèse de doctorat

Détection et suivi de personnes par vision omnidirectionnelle: approches 2D et 3D

Thèse de Doctorat de l'Université Paris-Saclay
Préparée à l'Université d'Evry Val d'Essonne et à
La Faculté des sciences de Rabat

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de Doctorat: Robotique

Thèse présentée et soutenue à Evry, le 14 mai 2018, par

Marouane Boui

Composition du Jury :

M. Fakhr-Eddine ABABSA — Directeur de thèse
Professeur, Arts et Métiers ParisTech

M. El Houssine BOUYAKHF — Directeur de thèse
Professeur, Faculté des Sciences, Université Mohammed V, Rabat

M. Hicham HADJ ABDELKADER — CoDirecteur de thèse
MCF, Université Evry Val d'Essonne

M. Pascal VASSEUR — Rapporteur
Professeur, Université de Rouen

M. El Hassan IBN ELHAJ — Rapporteur
Professeur,Institut National des Postes et Télécommunication (INPT) Rabat

M. Omar AIT AIDER — Examinateur
MCF, Université Clermont Auvergne

M. Mohammed Majid HIMMI — Examinateur
Professeur, Faculté des Sciences, (FSR), Université Mohammed V, Rabat

Mme Ibtissam BENMILOUD — Examinateur
Professeur, École Nationale Supérieure des Mines de Rabat (ENSMR)

Détection et suivi de personnes par vision omnidirectionnelle: approches 2D et 3D

M. BOUI

14 Mai 2018

Table des matières

Table des figures

Remerciements

Je tiens tout d'abord à remercier les directeurs de cette thèse :

Monsieur BOUYAKHF El Houssine qui fut pour moi un directeur de thèse attentif et disponible, malgré ses nombreuses charges. Sa compétence et sa rigueur scientifique, m'ont beaucoup aidé dans la réalisation de ce travail.

Monsieur ABABSA Fakhredine pour la confiance qu'il m'accordée en acceptant d'être mon directeur de thèse, pour ses multiples conseils et pour tout le temps qu'il a consacré à diriger cette recherche. De même que pour sa bienveillance et son attention particulière. J'ai beaucoup appris à ses côtés et je lui adresse ma gratitude et ma reconnaissance.

J'adresse de chaleureux remerciements à mon co-encadrant de thèse, Monsieur Hicham Hadj Abdelkader, pour la gentillesse et la patience qu'il a manifestées à mon égard durant cette thèse, ainsi que pour tous les conseils remarques qu'il m'a faites durant toute la durée de cette thèse. Son énergie et son franc-parler m'ont beaucoup aidé que ce soit sur le plan professionnel ou humain. J'ai pris un grand plaisir à travailler avec lui.

Je voudrais remercier les rapporteurs de cette thèse et notamment Monsieur Pascal Vasseur, Professeur des Universités à l'Université de Rouen, et Monsieur El Hassan IBN ELHAJ, Professeur de l'Enseignement Supérieur à l'Institut National des Postes et Télécommunication (INPT) de Rabat, pour l'intérêt qu'ils ont porté à mon travail. J'asso-

cie à ces remerciements Monsieur Omar Ait Aider, Professeur à l'Université Blaise-Pascal, Clermont, Madame Ibtissam Benmiloud, Professeur à l'École Nationale Supérieure des Mines de Rabat (ENSMR) et Monsieur Mohammed Majid Himmi, Professeur à la Faculté des Sciences de l'Université Mohammed V de Rabat, pour avoir accepté d'examiner mon travail.

Je tiens à remercier tous les membres de l'équipe IBISC (IRA2), pour leur aide et leur bonne humeur et leur soutien.

Je remercie également les amis qui m'ont permis de relativiser durant cette thèse. Merci de m'avoir permis de rester focaliser sur mon objectif.

Je remercie mon frère et mes sœurs, Hatim, Meriem et Laila, pour leur soutien et leur bonne humeur au cours de cette thèse. Je remercie également Samia Benchekroun pour la patience qu'elle a eue à mon égard durant la phase délicate de rédaction. Pour son soutien pendant mes périodes de doutes et pour ses multiples encouragements. Je voudrais lui dire simplement merci.

Mes plus profond remerciements vont a mon père qui m'a toujours soutenu. Je le remercierais jamais assez pour sa présence sans faille et son sens du dévouement. Tout au long de mon cursus, il ma toujours encouragé et aidé. Il a su me donner toutes les chances pour réussir. Qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de ses efforts ainsi que l'expression de ma plus affectueuse gratitude. Enfin, j'adresse toute mon affection à ma mère qui a toujours su trouver les mots justes pour me réconforter dans les nombreux moments de doute. Malgré mon éloignement depuis de nombreuses années sa confiance, sa tendresse et son amour, me portent et me guident tous les jours. Merci d'avoir fait de moi ce que je suis aujourd'hui.

À mes parents,

Résumé

Dans cette thèse, nous traiterons du problème de la détection et du suivi 3D de personnes dans des séquences d'images omnidirectionnelles, dans le but de réaliser des applications permettant l'estimation de pose 3D. Ceci nécessite, la mise en place d'un suivi stable et précis de la personne dans un environnement réel. Dans le cadre de cette étude, nous utiliserons une caméra catadioptrique composée d'un miroir sphérique et d'une caméra perspective. Ce type de capteur est couramment utilisé dans la vision par ordinateur et la robotique. Son principal avantage est son large champ de vision qui lui permet d'acquérir une vue à 360 degrés de la scène avec un seul capteur et en une seule image. Cependant, ce capteur va engendrer généralement des distorsions importantes dans les images, ne permettant pas une application directe des méthodes classiquement utilisées en vision perspective. Ce mémoire traite la description de deux approches de suivi développées durant cette thèse, qui permettent de tenir compte de ces distorsions. Elles illustrent le cheminement suivi par nos travaux durant ces trois années, nous permettant de passer de la détection de personnes à l'estimation 3D de sa pose. La première étape de nos travaux a consisté à mettre en place un algorithme de détection de personnes dans les images omnidirectionnelles. Nous avons proposé d'étendre l'approche conventionnelle pour la détection humaine dans une image perspective, basée sur l'Histogramme Orienté du Gradient (HOG), mise en place par Dallal

[Dalal, 2006], pour l'adapter à des images sphériques. Notre approche utilise les variétés riemanniennes afin d'adapter le calcul du gradient dans le cas des images omnidirectionnelles. Elle utilise aussi le gradient sphérique pour le cas des images sphériques afin de générer notre descripteur d'image omnidirectionnelle. Le descripteur sera utilisé avec un classifieur SVM pour la prise de décision. Plusieurs expériences ont été faites en utilisant la base de données d'images INRIA [Dalal, 2005], ainsi que notre base de données mise en place. Nous présenterons aussi les différents résultats obtenus avec notre algorithme pour une détection robuste de personnes dans les images omnidirectionnelles. Par la suite, nous nous sommes concentrés sur la mise en place d'un système de suivi 3D de personnes avec des caméras omnidirectionnelles. Nous avons fait le choix de faire du suivi 3D basé sur un modèle de la personne avec 32 degrés de liberté car nous nous sommes imposés comme contrainte l'utilisation d'une seule caméra catadioptrique. Notre travail a porté sur la mise en place de plusieurs fonctions de vraisemblance, basées sur les distances géodésiques dans l'espace sphérique SO_3, ainsi que sur la mise en correspondance de la silhouette dans l'image avec le modèle 3D projeté. Nos fonctions de vraisemblance combinées à un filtre particulaire (dont le modèle de propagation des particules est adapté à l'espace sphérique), permettent de faire un suivi 3D précis de la personne dans des images omnidirectionnelles. L'approche a été validée dans des conditions réelles et avec différents mouvements de la personne. Dans ce manuscrit, nous avons tenté de montrer que l'utilisation des caméras omnidirectionnelles dans le domaine de la détection et du suivi d'objets peut être précise si on tient bien compte des distorsions de ce type de capteur.

Mots clés : détection de personnes, caméra omnidirectionnelle, suivi monoculaire, estimation de pose.

Abstract

In order to realize applications allowing 3D pose estimation, we investigate the problem of 3D people detection and tracking in omnidirectional images sequences. This requires a stable and accurate monitoring of the person in a real environment. In order to achieve this, we will use a catadioptric camera composed of a spherical mirror and a perspective camera. This type of sensor is commonly used in computer vision and robotics. Its main advantage is its wide field of vision which allows it to acquire a 360-degree view of the scene with a single sensor and in a single image. However, this kind of sensor generally generates significant distortions in the images, not allowing a direct application of the methods conventionally used in perspective vision. Our thesis contains a description of two monitoring approaches that take into account these distortions. These methods show the progress of our work during these three years, allowing us to move from person detection to the 3D estimation of its pose.

The first step of this work consisted in setting up a person detection algorithm in the omnidirectional images. We proposed to extend the conventional approach for human detection in perspective image, based on the Gradient-Oriented Histogram (HOG) implemented by dallal [Dalal, 2006], in order to adjust it to spherical images. Our approach uses the Riemannian varieties to adapt the gradient calculation for omnidirectional images as

well as the spherical gradient for spherical images to generate our omnidirectional image descriptor. The descriptor will be used along with an SVM classifier for decision making. Several experiments have been done using the INRIA image database [Dalal, 2005], as well as our own database. We will introduce the different results obtained with our algorithm for a robust detection of people in omnidirectional images.

Subsequently, we set up a 3D tracking system for people with omnidirectional cameras. We have chosen to do a 3D tracking based on a model of the person with 32 degrees of freedom, because we have imposed as a constraint the use of a single catadioptric camera. Our work focused on the implementation of several likelihood functions, based on geodesic distances in the spherical space SO_3, as well as on the mapping of the silhouette in the image with the 3D model projected. Our likelihood functions combined with a particle filter (whose particle propagation model is adapted to spherical space) allows accurate 3D tracking of the person in omnidirectional images. The approach has been validated in real conditions and with different person moves. In this manuscript we have shown that the use of omnidirectional cameras in the field of object detection and tracking can be accurate if we take into account the distortions of this type of sensor.

Chapitre 1

Introduction

Sommaire

La vision par ordinateur est la discipline scientifique qui vise à extraire des informations à partir d'images acquises par ce dernier. L'un de ses objectifs, à long terme, est que les machines puissent percevoir ce qui les entoure, à l'image des humains, un rêve qui remonte aux années soixante-dix. L'analyse d'une image nous paraît (à l'Homme) triviale, vu que le cerveau humain peut quasi instantanément détecter une personne et la reconnaître. Pour les ordinateurs, la même image est uniquement une matrice de nombres. Cette complexité de la vision artificielle a été rapidement perçue et le problème a été divisé en plusieurs sous-problèmes plus simples. Dans cette thèse, nous nous sommes intéressés au problème de la détection et l'estimation de pose de la personne : ce qui va consister à localiser la personne dans l'espace. Nous avons démontré l'apport que peut avoir, dans ce domaine, l'utilisation de capteurs catadioptriques, ayant un champ de vision beaucoup plus important et cela

malgré la présence de distorsions dans l'image pouvant déconcerter l'œil humain dans un premier abord. La détection d'objets est un problème fondamental dans la vision par ordinateur. Elle constitue une condition préalable aux tâches de niveau supérieur, telles que la compréhension de scène ou la navigation autonome d'un robot. C'est aussi une étape cruciale pour de nombreuses applications de vision comme la surveillance visuelle, l'interaction homme-ordinateur, l'assistance au conducteur, l'annotation d'images automatiques ou la récupération d'images basées sur le contenu pour n'en citer que quelques-unes. La diversité des applications souligne l'importance de la détection des objets et explique les efforts immenses de la communauté scientifique dans ce domaine. Le défi de la détection d'objets est un défi majeur pour les ordinateurs. La tâche consiste à localiser les objets et à les annoter avec une étiquette pour cette classe d'objets. Les deux parties présentent de nombreuses difficultés. D'une part, un détecteur doit tenir compte des changements potentiels de la pose de l'objet tout en étant robuste face aux conditions d'éclairage, de pose et d'occlusion. D'autre part, les objets peuvent apparaître n'importe où dans une image. Cela produit un grand nombre d'emplacements d'objets possibles. Examiner ces cas d'une manière naïve est simplement impossible, notamment pour des applications en temps réel, mais aussi des plates-formes informatiques mobiles imposant des limites drastiques aux ressources informatiques disponibles. Par conséquent, l'efficacité est d'une importance primordiale pour que les systèmes soient pratiques. De plus, pour l'estimation, il faut définir un modèle 3D de la personne et trouver un compromis entre le nombre de poses possibles et le nombre de paramètres à estimer. En résumé, les détecteurs d'objets doivent gérer une énorme quantité de positions d'objets possibles et faire face à de grandes variations d'apparence, des exigences difficiles à satisfaire. Malgré les progrès significatifs réalisés au cours de la dernière décennie, la performance des détecteurs actuels reste encore loin de celle du système visuel humain.

1.1 Objectif

Cette thèse vise à mettre en place une approche robuste de suivi 3D de personnes capable de détecter et de suivre une personne dans diverses postures, sous différents points de vue, dans un environnement réel et dans un contexte de mono-caméras. La caméra est supposée fixe, elle observe la scène sous un seul point de vue. Pour atteindre cet objectif, un certain nombre de sous-objectifs ont été définis :

• Pour la partie détection, il faut identifier des caractéristiques importantes permettant de discriminer la classe personne. Ces caractéristiques doivent être insensibles aux conditions d'illumination et leur calcul adapté à un système temps-réel.

• Mettre en place une base de données conséquente d'images omnidirectionnelles permettant l'apprentissage et le test de la détection de personnes.

• Pour la partie suivi, il faut mettre en place le modèle 3D de la personne à utiliser ainsi que les fonctions de vraisemblance permettant le suivi 3D dans les images omnidirectionnelles. De plus, dans cette partie, il sera important de créer une base de données "vérité terrain" pour tester les différentes fonctions de vraisemblance mises en place.

1.2 Application

La détection et le suivi 3D de personnes dans les images et les vidéos est un domaine en plein essor. Il existe un grand nombre de domaines d'application, que ce soit pour le grand public ou pour les industriels. Ceci entraîne un grand nombre de recherches dans ce domaine, au vu des perspectives offertes. Dans cette section, nous présenterons quelques applications de la détection de personnes dans différents domaines :

Système de vidéo surveillance :

La détection de personnes dans les systèmes de vidéo surveillance est un sujet sur lequel beaucoup de recherches ont été menées [Paul et al., 2013]. D'ailleurs la plupart des caméras grand public de systèmes de surveillance intègrent désormais un algorithme de détection d'intrusion plus ou moins sophistiqué, car comme les caméras sont fixes ou ayant un mouvement prédéfini, la soustraction de l'arrière fond peut se faire rapidement. Aussi dans ces systèmes, on peut inclure la détection des chutes de personnes [Chua et al., 2015].

Système d'interaction homme-machine :

L'essor du dévellopement de la robotique humanoïde, mène à plus d'interaction entre l'homme et la machine. Ce qui implique, en amont un système de detection de personnes ou pour aller plus loin dans suivi 3D, permettant à l'homme d'interagir avec le robot, grâce notamment à la reconnaissance de gestes qui vont mener le robot à l'execution d'un ordre non verbal comme dans [Conseil, 2008]. De plus, avec la mise en place de caméra de type kinect, le processus permettant la reconnaissance du geste est facilité. Cela notamment grâce à l'intégration d'un système de capteur composé de deux caméras (RGB et profondeur), permettant la détection 3D d'une personne. Il faut rappeler que, initialement, la kinect à été développée pour les jeux vidéos avant d'être reprise dans le domaine de la recherche au vu de ses performances et de son champ d'application extrêmement vaste.

Détection de piétons :

Plusieurs systèmes d'assistance de conduite mis en place dans la dernière décennie sont liés à la sécurité des personnes, qui demeure un domaine prioritaire dans le cadre de la sécurité routière. C'est pourquoi, l'industrie automobile a investi beaucoup de fonds dans la recherche pour permettre la mise en place de voitures plus sûres pour le conduc-

teur mais aussi pour les piétons et les cyclistes. Les systèmes d'assistance de conduite permettent de détecter la présence d'un piéton sur la route et donc d'éviter la collision [Duchoň et al., 2012]. Comme par exemple, la mise en place par Volvo [Lindman et al., 2010] [Coelingh et al., 2010], dans son modèle Volvo S60, d'un système permettant la détection de piétons (figure 1.2) avec frein automatique en embarquant une caméra ainsi qu'un radar dans le véhicule. Le système de freinage automatique exige que l'objet soit vu à la fois par le radar et par la caméra. Le système peut détecter les piétons de 80 cm de hauteur et plus, ce qui permet la détection d'enfants. Néanmoins ce système a quelques limitations. Il ne peut pas détecter un piéton portant de gros objets. La capacité de la caméra à voir un piéton le matin à l'aube ou au crépuscule est limitée. La fonction de détection de piétons est désactivée dans l'obscurité ou dans les tunnels, même si la zone dispose d'éclairage publique. D'autres systèmes d'assistance peuvent être cités comme le système EyeSight de Subaru, disponible au Japon et similaire à celui de Volvo, dont la dernière version repose sur deux caméras pour produire une vision stéréo, au lieu d'un radar et d'une seule caméra.

1.3 Organisation de la thèse

La structure de la thèse est la suivante. Le chapitre 2 formalise la théorie utilisée pour la formation des images omnidirectionnelles avec différents types de capteurs. Elle présente le modèle unifié que nous utiliserons dans nos travaux. Ce chapitre traite en outre de la calibration de ce type de capteurs. Le chapitre 3 fournit une première implémentation qui s'appuie sur la recherche du gradient dans l'espace sphérique et omnidirectionnel. Le modèle permet de faire une détection 2D de la personne. Il est basé sur l'utilisation d'un HOG adapté aux images omnidirectionnelles, combiné avec un SVM. Nous présentons aussi les résultats de détection comparés à la méthode de HOG classique. Le chapitre 4

FIGURE 1.1 – Un exemple de la façon dont les piétons sont détectés dans un système d'assistance de conduite [Lindman et al., 2010]

présente nos travaux dans la détection et le suivi 3D de personnes. Nous utiliserons la méthode présentée dans le chapitre précédant comme initialisation à notre algorithme de suivi 3D. Cet algorithme sera basé sur un modèle 3D de la personne avec 30 degrés de liberté, combiné avec un filtre particulaire. Nous présenterons dans ce chapitre différentes fonctions de vraisemblance adaptées à l'espace sphérique. Elles seront basées sur les distances géodésiques. Le chapitre 5 conclut cette thèse et fournit les différentes perspectives permettant d'approfondir cette recherche.

Chapitre 2

Vision Omnidirectionnelle

Sommaire

Dans ce chapitre, nous donnons un bref aperçu des systèmes proposés pour créer des caméras omnidirectionnelles dites centrales, c'est-à-dire ayant un point de vue unique. Nous examinons ces modèles pour les configurations catadioptrique, dioptrique et poly-dioptrique. Enfin, nous présenterons le modèle unifié ; en combinant une sphère virtuelle

avec un modèle sténopé (pinhole) qui permet de modéliser le processus de formation d'images pour les systèmes omnidirectionnels, ainsi que pour les caméras perspectives.

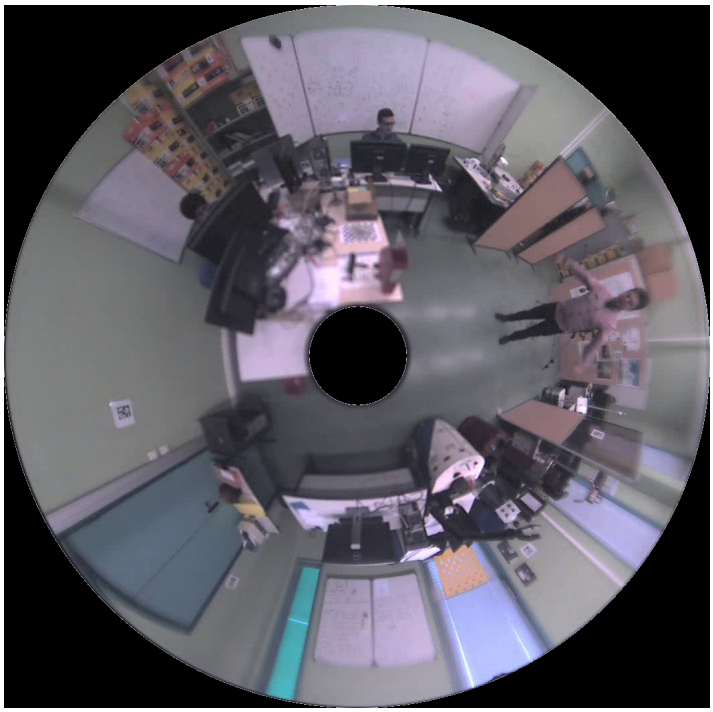


FIGURE 2.1 – Exemple d'une image omnidirectionnelle.

2.1 Capteur omnidirectionnel

Une image panoramique correspond à une observation de 360° autour du capteur. Comme montré sur la figure 2.1, il faudrait une série d'images prises à plusieurs angles de vue pour obtenir cette quantité d'informations avec une seule caméra conventionnelle. Pour concevoir une caméra panoramique, trois possibilités majeures nous sont offertes : un ensemble de systèmes multi-caméras pour créer une mosaïque, une caméra conventionnelle

avec un objectif Fisheye. La troisième solution consiste à combiner un miroir avec une caméra conventionnelle. Nous montrerons également que l'utilisation d'une caméra omnidirectionnelle crée des distorsions, lors du passage de l'image omnidirectionnelle à l'image perspective, ce qui rend l'utilisation inadéquate des algorithmes en vision classique comme par exemple la détection de personnes dans une image omnidirectionnelle.

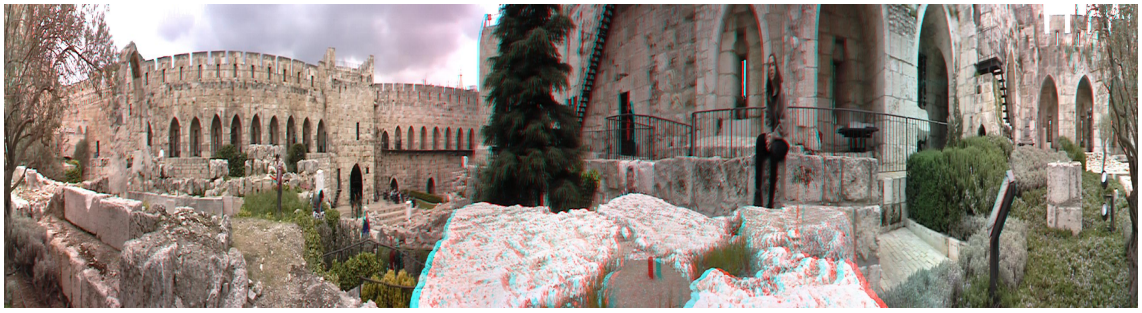
(a) Dodeca (b) mini eye 4 (c) Ladybug3

FIGURE 2.2 – Capteurs Poly-dioptriques

2.1.1 Capteur poly-dioptrique

Les capteurs Poly-dioptriques peuvent être créés grâce à un réseau de capteurs (Figure 2.2). Cette méthode présente l'avantage d'avoir une haute résolution sur tout le champ de vision du système. Le large champ de vision est obtenu en fusionnant les images provenant de plusieurs caméras. Les contraintes mécaniques entraînent des désalignements optiques entre chaque caméra de sorte que les différentes caméras composant le réseau peuvent ne pas partager le même centre de projection. Dans cette situation, les images avec un objet capturé à de faibles distances peuvent ne pas se fondre, et des artefacts visuels peuvent

apparaître. Pour être opérationnels, les capteurs poly-dioptriques doivent être calibrés avec une grande précision. De plus, un dernier inconvénient majeur de ce type de capteurs est la nécessité de la synchronisation des capteurs composant le système pour éviter les artefacts lors du mélange d'images. La figure 2.2 présente un panel de capteurs poly-dioptriques. Dans cette partie on peut aussi citer la reconstruction d'images panoramiques



(a)

(b)

FIGURE 2.3 – (a) Image omnidirectionnelle obtenue avec une seule caméra aprés plusieurs rotations (b) Caméra perspective sur un axe de rotation [Peleg et al., 2001b]

qui peuvent être obtenues en utilisant plusieurs images prises à partir d'une seule caméra qui effectue une rotation autour d'un axe donné. Dans [Peleg et al., 2001a], Pritch met en place une caméra "OmniStereo", lui permettant de créer des images panoramiques, qui peuvent être obtenues à l'aide d'images capturées avec une caméra régulière tournant autour d'un axe. Le collage des différentes images ainsi prises, permet de générer une image

panoramique de forme cylindrique comme on peut le voir dans la (figure 2.3). Avec ces systèmes, l'acquisition et l'association des données sont rarement en temps réel. De plus, cette caméra présente un autre inconvénient lors de l'acquisition des scènes en mouvement. En effet, si une personne change de place lors des prises de vue consécutives, elle risque d'apparaître plusieurs fois sur l'image panoramique.

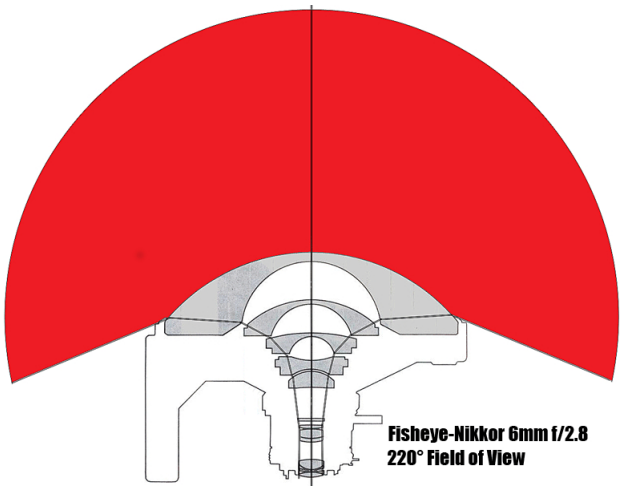
2.1.2 Capteur dioptrique

Les capteurs dioptriques sont des systèmes qui n'utilisent que des optiques réfractives, comme la lentille fish-eyes. Cet objectif présente l'avantage de se fixer sur les caméras standards à la place des objectifs classiques (figure 2.4). De plus, il permet de ne pas occulter une partie du champ visuel contrairement aux systèmes présentés dans la section suivante. Cependant, ces capteurs ne vérifient pas exactement la contrainte de point de vue unique. En plus, une limitation directe de ce type de capteur est la distorsion de l'image finale qui se traduit par une définition angulaire non uniforme (les bordures ont beaucoup moins de pixels par degré que le centre).

2.1.3 Capteur catadioptrique

Le terme catadioptrique provient de la contraction de catoptrique (qui fait référence à la surface de réflexion comme les miroirs) et de dioptrique, pour la vision par réfraction. Un capteur catadioptrique peut être obtenu en plaçant une caméra en perspective devant un ou plusieurs miroirs convexes. Nous pouvons ainsi obtenir une vue à 360° de l'environnement. L'acquisition est en temps réel avec une bonne résolution au centre de l'image (figure 2.5). Deux types de miroir vérifient la contrainte de point unique, à savoir les miroirs paraboloïde et hyperboloïde. Le principal inconvénient de ces systèmes est la résolution non uniforme et la faible résolution par rapport au système multi-caméras. Cette faible résolution est due au

(a)



(b)

FIGURE 2.4 – Capteur lentille fisheyes [Ardouin, 2014]

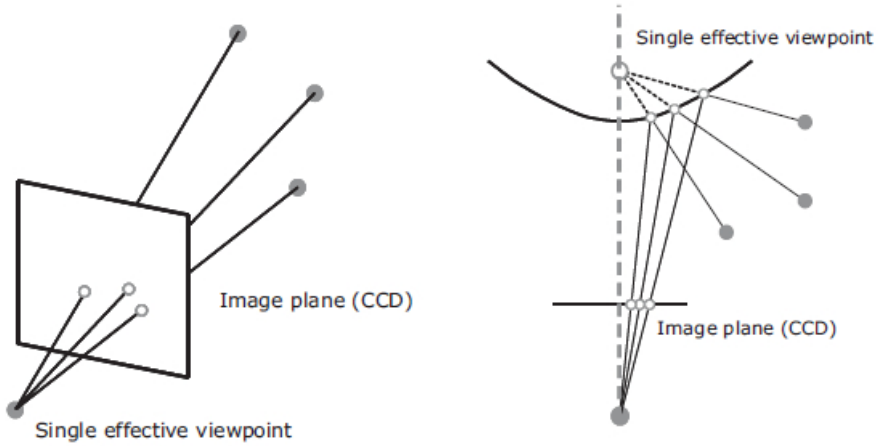
nombre important de données contenues en une seule image. Malgré ces inconvénients, les caméras catadioptriques deviennent un choix populaire pour les applications en robotique. En effet, elles permettent un élargissement du champ de vue avec pour unique contrepartie

la génération de distorsion dans l'image. Dans le cadre de cette thèse, nous utiliserons principalement les caméras catadioptriques.

FIGURE 2.5 – Différents types de miroirs composant un capteur catadioptrique

2.1.4 Formation des images catadioptriques

La contrainte de point de vue unique correspond au fait qu'un point 2D dans l'image ne mesure l'intensité que d'un unique rayon lumineux dans l'espace 3D. L'ensemble des rayons reflétés par le miroir de révolution, s'intersectant en un point unique, le centre de projection du miroir. Les capteurs respectant cette contrainte sont dits à projection centrale ou capteurs centraux. Par opposition, ceux possédant plusieurs centres de projection sont dits à projection non-centrale ou capteurs non-centraux. L'exemple classique d'un capteur central est le système d'acquisition perspectif dont la transformation entre l'espace 3D et le plan image est décrite par le modèle « pinhole ». Dans le cas des capteurs catadioptriques,



(a) (b)

FIGURE 2.6 – Système vérifiant la condition de point de vue unique. (a) dans le cas d'une caméra perceptive. (b) dans le cas d'une caméra catadioptrique avec un miroir hyperbolique

la transformation entre l'espace 3D et le plan image ne peut être décrite uniquement avec le modèle « pinhole ». Les distorsions introduites par le miroir étant importantes. Néanmoins, pour certaines configurations miroir-caméra, le capteur catadioptrique peut être considéré comme central. La contrainte de point de vue unique est alors définie comme la condition, que l'ensemble des rayons provenant du point focal « effective pinhole », intersectant le miroir, semble converger en un point (figure 2.6). L'intérêt de la contrainte de point de vue unique est essentiel, parce qu'elle permet l'utilisation des résultats établis en vision classique, ce qui conduit à une simplification des modèles de projection. L'image acquise peut être rétro-projetée sur tout type de surfaces (projection cylindrique, sphérique, etc.). La géométrie épipolaire permet la recherche de points homologues pour des systèmes stéréoscopiques. Baker et Nayar [Baker and Nayar, 1998], ont établi la classe de tous les capteurs catadioptriques centraux. formulation du problème de respect de la contrainte

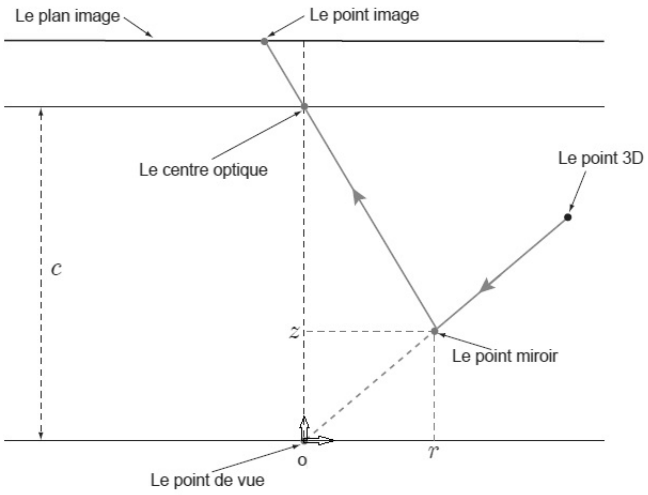


FIGURE 2.7 – Formation d'une image catadioptrique à point de vue unique [Baker and Nayar, 1999]

d'unicité du point de vue dans les cas des capteurs catadioptriques à projection centrale. Les rayons réfléchis sur le miroir doivent se prolonger en un point virtuel appelé point de vue effectif, ici placé à l'origine du repère. L'équation paramétrée du miroir peut être obtenue en considérant la loi de réflexion liant les angles incident et réfléchi et la normale à la surface. La contrainte de point du vue peut être exprimée en considérant un miroir ponctuel dans le plan cartésien (O, r, z), en posant $r = \sqrt{x^2 + y^2}$. On aboutit alors à des familles de miroirs paramétrées par les constantes c et k et décrites par les équations :

$$\left(z - \frac{c}{2}\right)^2 - r^2\left(\frac{k}{2} - 1\right) = \frac{c^2}{4}\left(\frac{k-2}{k}\right) \quad (k \geq 2) \tag{2.1}$$

$$\left(z - \frac{c}{2}\right)^2 + r^2\left(1 + \frac{c^2}{2k}\right) = \left(\frac{2k + c^2}{4}\right) \quad (k > 0) \tag{2.2}$$

avec et c la distance entre le point de vue et le centre optique de la caméra et k, une constante strictement positive (voir figure 2.7). Dans la première de ces deux équations, le paramètre constant k est contraint par $k > 2$ (plutôt que $k > 0$) puisque $0 < k < 2$ conduit à des solutions complexes. Les miroirs satisfaisant la contrainte du point de vue unique sont : le miroir plan, le miroir elliptique, le miroir parabolique et le miroir hyperbolique.

Le miroir plan : Si on pose $k = 2$ et $c > 0$ dans l'équation (2.2) , nous obtenant alors :

$$z = \frac{c}{2} \tag{2.3}$$

La figure 2.8 montre la configuration d'un capteur catadioptrique à base d'un miroir plan.

(a) (b)

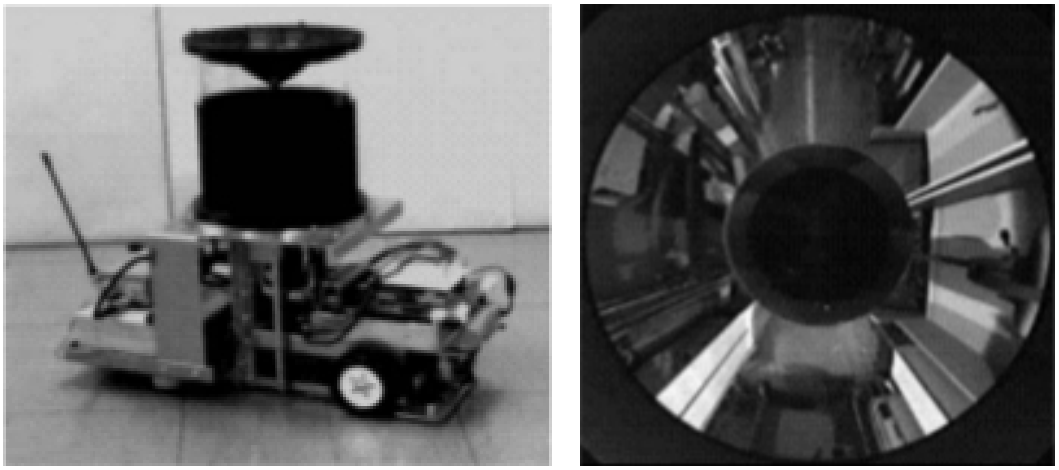
FIGURE 2.8 – Miroir plan avec multiples caméras. (a) FullView Nalwa (www.fullview.com). (b) Microsoft roundtable

Ce qui représente l'equation d'un plan. Ainsi le miroir plan est placé en parallèle au

plan image sur la médiatrice du segment joignant le point de vue et le centre optique de la caméra. Même si cette configuration satisfait la contrainte du point de vue unique, elle ne permet pas pour autant d'accroître le champ de vision. Pour remédier à ce problème, Nalwa [Nalwa, 1996] a conçu un système à base de quatre miroirs et de quatre caméras disposés en pyramide . Ce système est facile à implémenter et permet de générer une image omnidirectionnelle à haute résolution de 360°x50°. Gluckman et Nayar ont présenté dans [Gluckman 1999] un système à base de deux miroirs plans placés en face d'une caméra. L'image panoramique est produite par une mise en correspondance stéréoscopique.

Le miroir conique : dans l'équation (2.2), si on pose $k >= 2$ et $c = 0$, alors on a :

$$z = \pm r\sqrt{\frac{k}{2} - 1} \tag{2.4}$$



(a) (b)

FIGURE 2.9 – Capteur omnidirectionnel avec un miroir conique. (a) COPIS. (b) Image acquise avec la caméra COPIS

Ce sera l'équation d'une conique, et puisque c = 0, le centre optique doit être positionné sur le sommet du cône. De tous les rayons lumineux en provenance de la scène, seuls ceux tangents à la surface du cône seront captés par la caméra. Par conséquent, cette solution ne peut pas être employée pour élargir le champ de vision. Ainsi le miroir conique est considéré comme une solution dégénérée. Toutefois, il est possible de mettre en place un système avec des miroirs coniques, mais qui ne respecte pas la condition de point de vue unique. Par exemple, les travaux de Yagi et Kawato [Yagi and Kawato, 1990] ont mis en place un capteur catadioptrique non central appelé COPIS (COnic Projection Image Sensor). Ce système à base d'un miroir conique a été utilisé dans la navigation d'un robot mobile évoluant dans un environnement réel.

Le miroir sphérique : dans l'équation (2.1), si on pose $k > 0$ et $c = 0$, alors on a :

$$z^2 + r^2 = \frac{k}{2} \tag{2.5}$$

Le miroir sphérique est une autre solution dégénérée des capteurs catadioptriques centraux. En effet, pour que le miroir sphérique respecte la contrainte du point de vue unique, il faut que le centre optique, le point de vue unique et le centre de la sphère soient confondus. Dans ce cas de figure, l'image obetenue est la caméra elle même (figure 2.10). Néanmoins, les travaux de Fial et Alta [Fiala and Basu, 2003], [Fiala and Basu, 2005] ont utilisé des capteurs omnidirectionnels avec des miroirs sphériques, en ne considérant pas le système comme à point de vue unique.

Le miroir paraboloïde : dans l'équation (2.1), si on pose k infini et c infini, alors on a :

$$z = \frac{h^2 - r^2}{2h} \tag{2.6}$$

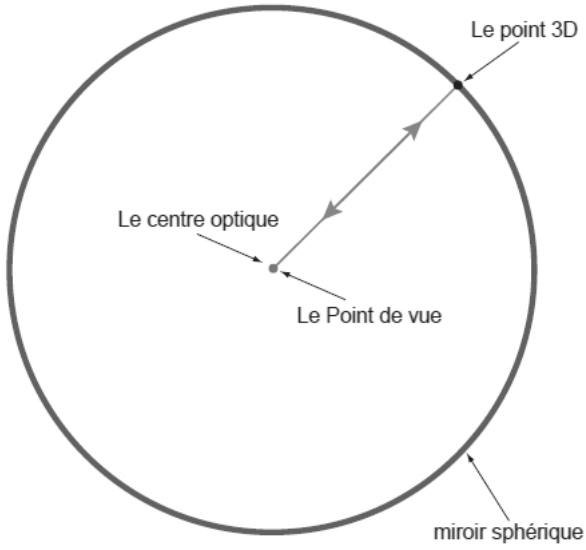


FIGURE 2.10 – Système catadioptrique à miroir sphérique

avec $h = \frac{c}{k}$.

La forme parabolique est une solution de la contrainte du point de vue unique, dans le cas où le point de vue unique se trouve à une distance infinie, ce qui correspond à la projection orthographique. Le point de vue coïncide avec le foyer du miroir et les rayons incidents sont réfléchis parallèlement à l'axe de symétrie du miroir. Par conséquent, les miroirs paraboliques sont combinés avec des caméras orthographiques (figure 2.11). Il est aussi possible de combiner un miroir parabolique avec une lentille afin de l'utiliser avec une caméra perspective. Ce montage est alors simple à réaliser et assure la conception d'un capteur catadioptrique idéal. Cependant, les principaux inconvénients d'un tel montage, sont la taille de la lentille télécentrique et son coût.

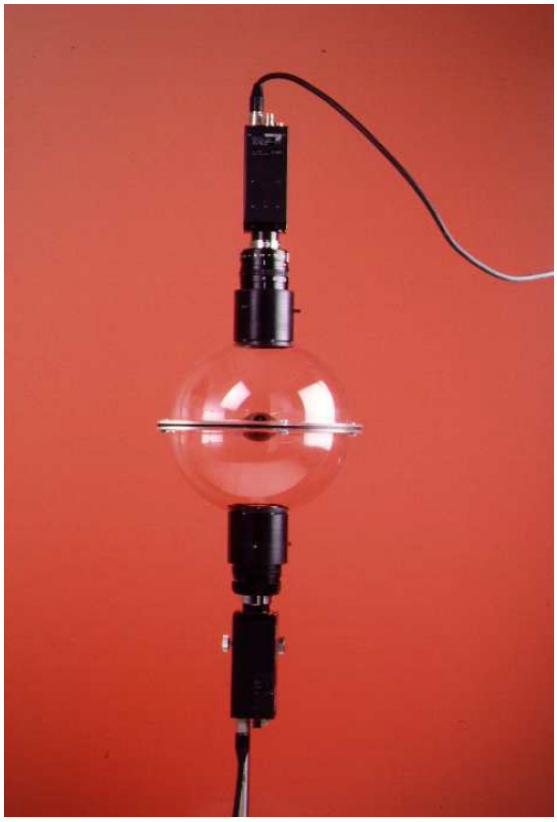


FIGURE 2.11 – Système catadioptrique composé de deux caméras à miroir paraboloïde, fournissant un champ de vision équivalent à une sphère.

Le miroir hyperboloïde : dans l'équation (2.1), si on pose $k > 2$ et $c > 0$ alors on a :

$$\frac{1}{a^2}(z - \frac{c}{e} - \frac{1}{b^2}) = 1 \tag{2.7}$$

avec :

$$a = \frac{c}{2}\sqrt{\frac{k-2}{k}} \quad \text{et} \quad b = \frac{c}{2}\sqrt{\frac{2}{k}}$$

le miroir hyperbolique satisfait la contrainte du point de vue unique lorsque le centre optique de la caméra perspective coïncide avec le deuxième foyer de l'hyperbole (figure

FIGURE 2.12 – Système catadioptrique composé d'une caméra et d'un miroir hyperbolique.

2.12). La caméra permet d'avoir un champ de vision plus important, si le paramètre k est important. Néanmoins, la construction de ce type de caméra nécessite une très grande précision pour satisfaire la contrainte du point de vue unique, car le centre optique de la caméra doit coïncider avec le deuxième foyer de l'hyperboloïde.

Le miroir elliptique : dans l'équation (2.1), si on pose $k > 0$ et $c > 0$ alors on a :

$$\frac{1}{a^2}(z - \frac{c^2}{e} + \frac{1}{b^2}r^2) = 1 \tag{2.8}$$

avec :

$$a = \sqrt{\frac{2k + c^2}{4}} \quad \text{et} \quad b = \sqrt{\frac{k}{2}}$$

Le point de vue et le centre optique de la caméra perspective coïncident avec les deux

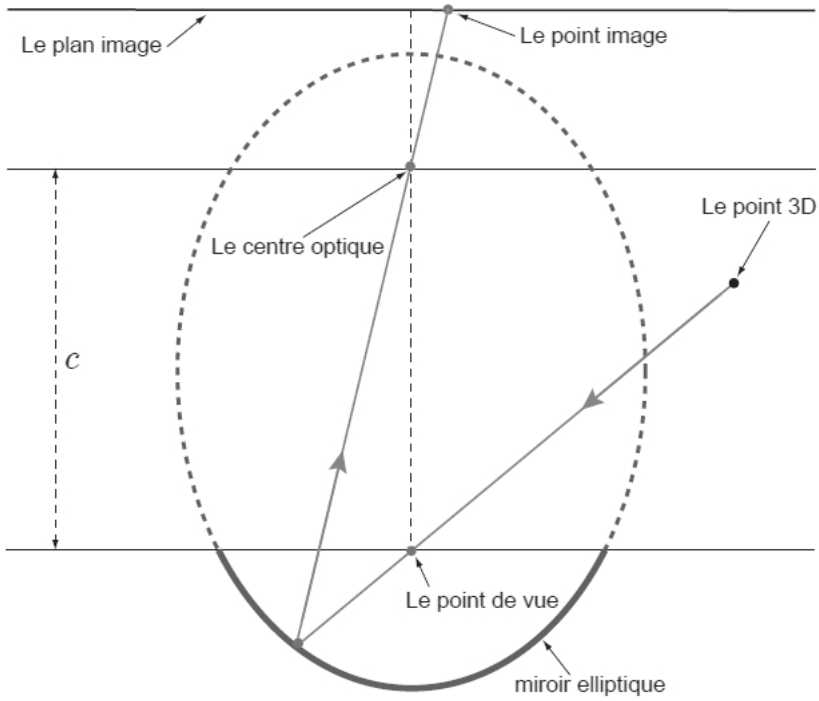


FIGURE 2.13 – Formation de l'image dans un système catadioptrique à miroir elliptique, fournissant un champ de vision équivalent à une demi-sphère

foyers du miroir. Ce type de miroirs est dénué d'intérêt pratique. En effet, le champ de vision d'un miroir elliptique est limité à cause de sa forme concave (figure 2.13). En 1998, Nene et Nayar ont proposé un système stéréoscopique catadioptrique à base de deux miroirs elliptiques [Nene and Nayar, 1998]. Le système est configuré de telle sorte, que l'un des foyers de chacun des miroirs coïncide avec le centre optique (figure 2.14). Les rayons lumineux réfléchis par les miroirs passent par les points de vue v1 et v2 situés sur les deux autres foyers des miroirs.

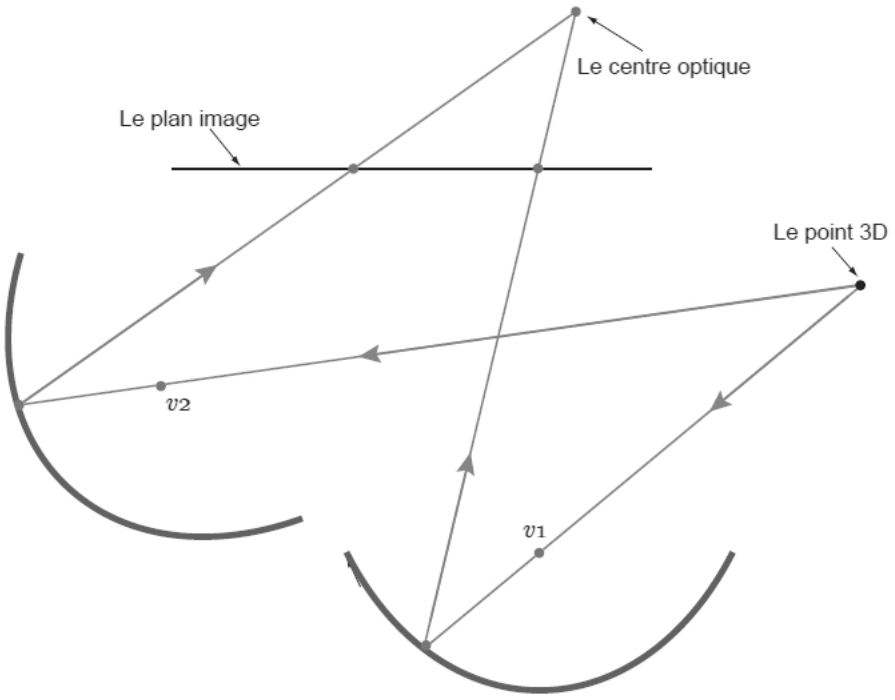


FIGURE 2.14 – Formation de l'image catadioptrique à base de deux miroirs elliptiques de Nene

2.2 Modélisation des caméras centrales

2.2.1 Modèle perspectif

Dans cette section nous présenterons le modèle de projection sténopé ("pinhole" ou trou d'épingle). Le modèle de caméra sténopé a été établi il y a longtemps. Leonardi Da Vinci l'a popularisé au 15ème siècle. La figure 2.15 correspond à un dessin réalisé par Reinerus Gemma-Frisius en 1544. Reinerus est un mathématicien et médecin néerlandais, qui a observé une éclipse du soleil avec une "caméra obscura" à Louvain le 24 janvier 1544. Un an plus tard, il fait une illustration de l'événement dans son livre De Radio Astronomica

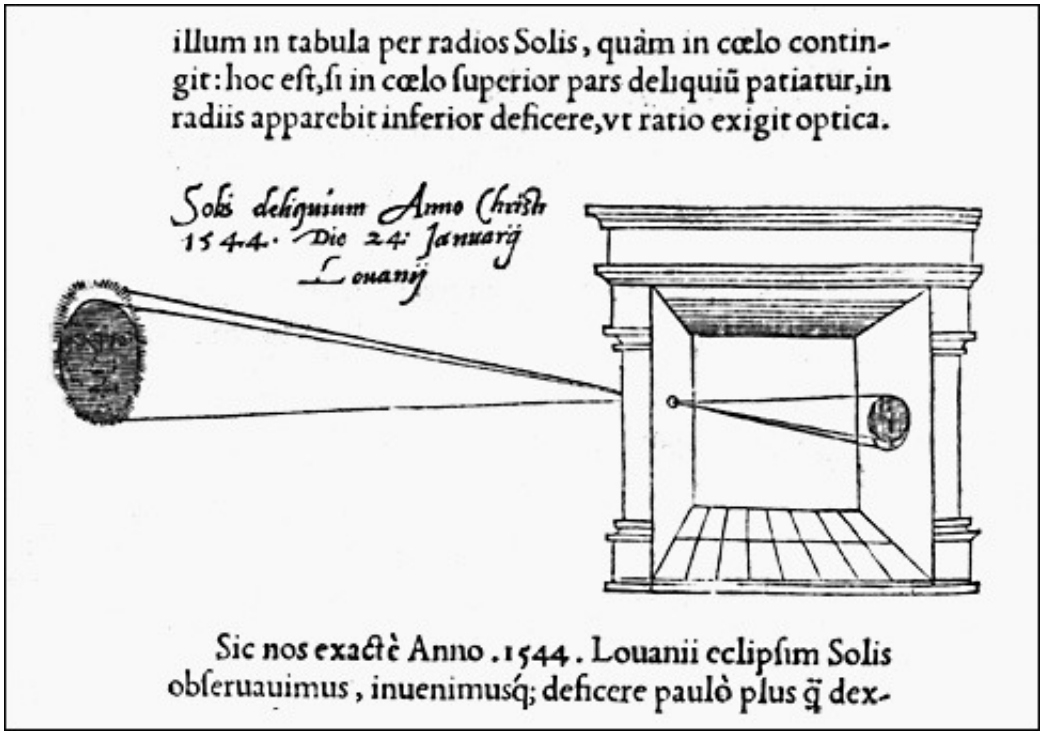


FIGURE 2.15 – L'illustration de la "camera obscura" par Frisieus Gemmas en 1544

et Geometrica (figure 2.15). C'est la première publication d'une illustration d'une "caméra obscura" et c'est une excellente illustration de la projection sténopé. L'image inversée du soleil et de la lune est clairement visible sur le mur à droite. De nos jours, les caméras conventionnelles sont encore représentées avec ce modèle qui reste encore largement utilisé, car il permet de modéliser de manière simple une projection perspective entre l'espace 3D et le plan image. Soit C le centre optique d'une caméra perspective et est l'origine du repère lié à cette dernière (figure 2.16). Soit I le plan image, situé à une distance f (la distance focale de la caméra) du centre de la caméra. La projection d'un point Q de l'espace monde est définie comme étant l'intersection q du rayon de projection (CQ) avec le plan image. Cette projection peut être représentée par une transformation projec-

tive M de P^3 vers P^2. La matrice de projection M ne fournit pas d'indications sur les

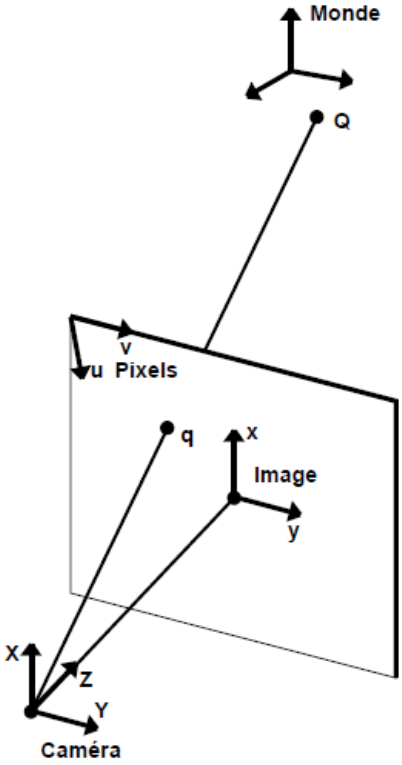


FIGURE 2.16 – Le modèle sténopé

grandeurs physiques mises en jeu durant le processus de prises d'images. Il est donc judicieux de décomposer M sous forme d'un produit matriciel faisant clairement apparaître les différents paramètres intervenant durant le processus de formation des images. Mais avant, nous introduisons quelques repères de coordonnées par rapport auxquels nous allons définir les transformations composant M.

- Un repère global par rapport auquel sont donnés les points tridimensionnels, nous l'appellerons le repère monde.

- Un repère 3D local attaché à la caméra, le repère caméra.
- Un repère 2D du plan image, le repère image.
- Un deuxième repère 2D du plan image, le repère pixel, associé à la grille des pixels.

Ainsi nous pourrons décomposer M pour faire apparaître, d'une part, les paramètres extrinsèques qui expriment la relation de passage du repère monde au repère caméra. D'autre part, les paramètres intrinsèques décrivent les caractéristiques propres au système d'acquisition. Ils permettent le passage du repère image métrique au repère pixélique.

L'équation de projection d'un point de l'espace 3D en coordonnées homogènes $Q = (X; Y; Z; 1)^T$ définies dans le repère monde et d'un point 2D en coordonnées homogènes $q = (u; v; 1)^T$ définies dans le repère pixelique, est donnée par l'équation (2.9). Cette équation fait apparaître une matrice homogène de changement de base $S = (R|\mathbf{t})$, modélisant les paramètres extrinsèques, avec t la translation entre le repère monde et le repère caméra et R la matrice de rotation (une matrice orthogonale). R indique l'orientation de la caméra par rapport au monde et t sa position. La matrice homogène $\mathbf{P}_I$ est la projection perspective proprement dite, permettant à un point $Q = (X; Y; Z; 1)^T$ dans le repère caméra de se projeter sur un point $\mathbf{x} = (x; y; 1)^T$. La matrice homogène des paramètres intrinsèques K pour une projection perspective dépend de la focale f, du nombre de pixels par unité de longueur, sur chacun des axes avec le facteur d'échelle vertical k_u (pixels/m), le facteur d'échelle horizontale k_v et les coordonnées du centre de l'image $c(u_0, v_0)$ dans le repère pixels :

$$x\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \overbrace{\overbrace{\begin{pmatrix} k_u f & 0 & u_0 \\ 0 & k_v f & v_0 \end{pmatrix}}^{K}\overbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}^{P_I}}^{}\overbrace{\begin{pmatrix} R & t \\ 0_{3\times 1}^T & 1 \end{pmatrix}}^{S}\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2.9}$$

2.2.2 Modèle unifié

Dans cette section, nous rappelons le model adopté qui est la projection centrale unifiée en utilisant la sphère unitaire. Geyer [Geyer and Daniilidis, 2000] et Barreto [Barreto and Araujo, 2001] ont proposé un modèle générique de formation d'images omni-directionnelles qui dissocie les non-linéarités introduites par le miroir. Il s'agit du modèle de la sphère d'équivalence qui se base sur une représentation sphérique de l'image. Pour nos travaux nous avons choisi d'utiliser le modèle introduit par [Mei and Rives, 2007] où la projection d'un point 3D se fait selon les étapes suivantes :

1- Projection d'un point de l'espace dans le repère miroir sur la sphère

$$(\chi)_{\mathcal{F}_m} \rightarrow (\chi_s)_{\mathcal{F}_m} = \frac{\chi}{\|\chi\|} = (X_s, Y_s, Z_s) \tag{2.10}$$

2- Changement du centre du repère $C_m \longrightarrow C_p$

$$(\chi_s)_{\mathcal{F}_m} \rightarrow (\chi_s)_{\mathcal{F}_P} = (X_s, Y_s, Z_s + \xi) \tag{2.11}$$

3- Normalisation :

$$m = \left(\frac{X_s}{Z_s + \xi}, \frac{Y_s}{Z_s + \xi}, 1\right) = h(\chi_s) \tag{2.12}$$

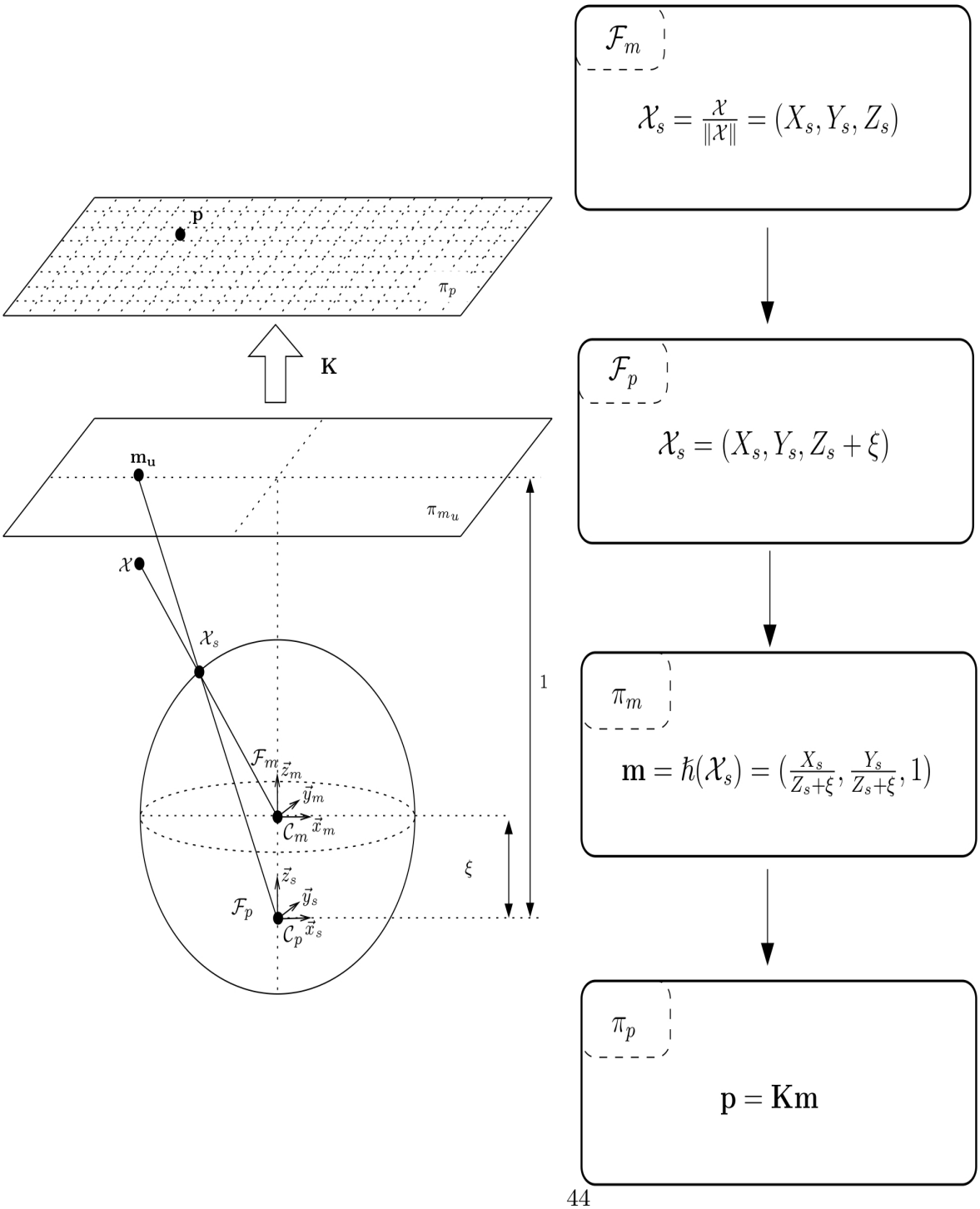


FIGURE 2.17 – Le modèle de projection unifié [Mei and Rives, 2007].

4- Passage du plan normalisé vers le plan image pixélique :

$$\mathbb{P} = Km = \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} m = k\left(m\right) \tag{2.13}$$

Avec f_u et f_v qui expriment la distance focale, en nombre de pixels en direction horizontal et vertical. , (u_0, v_0) le point principal. Les pixels sont considérés carré ou rectangle. La projection inverse peut être utilisée pour représenter une image omnidirectionnelle en une image sphérique. Comme le montre Barreto dans [Barreto et al., 2003] :

$$h^{-1}\left(m\right) = \begin{bmatrix} \frac{\xi+\sqrt{1+(1-\xi^2)(x^2y^2)}}{x^2+y^2+1}x \\ \frac{\xi+\sqrt{1+(1-\xi^2)(x^2y^2)}}{x^2+y^2+1}y \\ \frac{\xi+\sqrt{1+(1-\xi^2)(x^2y^2)}}{x^2+y^2+1} - \xi \end{bmatrix} \tag{2.14}$$

$$h^{-1}\left(m\right) \sim \begin{bmatrix} x \\ y \\ 1 - \xi\frac{x^2+y^2+1}{\xi+\sqrt{1+(\mathbb{I}-\xi^2)(x^2+y^2)}} \end{bmatrix} \tag{2.15}$$

Ce modèle de projection unifié est valide pour toute caméra à point de vue unique, y compris les caméras perspectives. En effet, il suffit de prendre $\xi = 0$, pour retrouver une projection perspective (modèle sténopé). ξ est d'ailleurs le paramètre dont la valeur dépend de la forme du miroir. Enfin, même si la caméra fisheye n'est pas à point de vue unique, ce modèle de projection est une bonne approximation de la réelle projection et peut être utilisé pour modéliser certaines caméras de ce type [Ying and Hu, 2004]. Le modèle introduit par [Mei, 2007] permet d'ajouter une distorsion radiale et tangentielle dans le modèle de formation des images par rapport à d'autres modèles présents dans la littérature. Son modèle permet de rendre les paramètres de calibrage facilement identifiables et présente

un bon compromis entre un modèle trop générique et une sur-paramétrisation.

2.2.3 Modèle de Taylor

Dans cette section, nous présentons un autre formalisme introduit par Micusik [Micušık, 2004]. Après avoir présenté le modèle unifié de Geyer et Daniilidis pour les caméras catadioptriques centrales, nous examinons brièvement celui proposé par Micusik pour les caméras fisheye. Par la suite, nous présenterons le modèle de Taylor introduit par Scaramuzza [Scaramuzza et al., 2006a] [Scaramuzza and Siegwart, 2007] qui convient pour les caméras omnidirectionnelles centrales dioptriques et catadioptriques. La nouveauté de cette approche par rapport aux travaux antérieurs est qu'elle décrit le modèle de formation des images de la caméra en termes d'expression polynomiale dont les coefficients seront les paramètres d'étalonnage du capteur. De plus, Scaramuzza et al proposent une nouvelle méthode permettant son étalonnage en utilisant le modèle polynomial [Scaramuzza et al., 2006b].

Ainsi, selon le formalisme de Micusik [Micušık, 2004], les points d'image des caméras omnidirectionnelles peuvent être représentés dans un modèle sphérique comme un ensemble de vecteurs unitaires dans R^3, de sorte qu'un vecteur correspond juste à l'un des sous-espaces demi-dimensionnels de R^3. Cela signifie qu'un point d'image représente tous les points de la scène situés sur une demi-ligne émanant du centre de la caméra contrairement au modèle perspectif, où un point d'image représente tous les points de la scène situés sur toute la ligne passant par le centre optique. Ainsi, l'équation de projection pour les caméras omnidirectionnelles peut être écrite comme suit :

$$\lambda q = P.X, \lambda > 0, \tag{2.16}$$

où $q = [x, y, z]$ est un vecteur unitaire (c'est-à-dire $\|q\| = 1$) représentant le point d'image. En ce qui concerne les caméras omnidirectionnelles, les hypothèses suivantes sont toujours prises en considération :

1. Le miroir (ou la lentille fisheye pour les caméras dioptriques) est symétrique en rotation par rapport à son axe. Cette symétrie est garantie par le fabriquant.

2. L'axe du miroir (lentille) est perpendiculaire au plan du capteur.

Dans ce qui suit, nous allons considérer des caméras omnidirectionnelles centrales et donc à point de vue unique. Supposons que nous observons un point 3D X par une caméra omnidirectionnelle (figure 2.19). En utilisant le modèle sphérique donné par l'équation 2.16, il existe toujours un vecteur $p'' = (x''^T, z'')$, ayant la même direction que q, qui est projeté au point u'' sur le capteur plan, de sorte que u'' soit colinéaire avec x''. Cela peut être formalisé comme suit :

$$p'' = \begin{bmatrix} h(\|u''\|)u'' \\ g(\|u''\|) \end{bmatrix}, \tag{2.17}$$

Avec h et g deux fonctions de $R \longrightarrow R$, qui dépendent de la distance $\|u''\|$ du point u'' à l'origine de l'axe du capteur (centre de symétrie). Les fonctions h, g sont différentes pour différents types de lentilles et miroirs. Pour les lentilles fisheyes, ces fonctions dépendent du type de lentille (par exemple, equisolide, equiangulaire, etc.) et pour les miroirs, elles dépendent de la forme du miroir (par exemple parabolique, hyperbolique). La projection du vecteur p'' sur le plan du capteur donne le point u'' grâce aux fonctions h et g. La figure 2.19 représente cette projection pour une lentille fisheye et un miroir hyperbolique. Dans le cas des lentilles fisheyes, nous avons h = 1 et donc le vecteur u'' est une projection orthographique du point p'' sur le plan du capteur. Dans le cas des miroirs, le vecteur p'' est projeté à travers une caméra perspective ayant le centre optique C (le centre optique peut se trouver à l'infini pour les miroirs paraboliques). Nous pouvons également constater

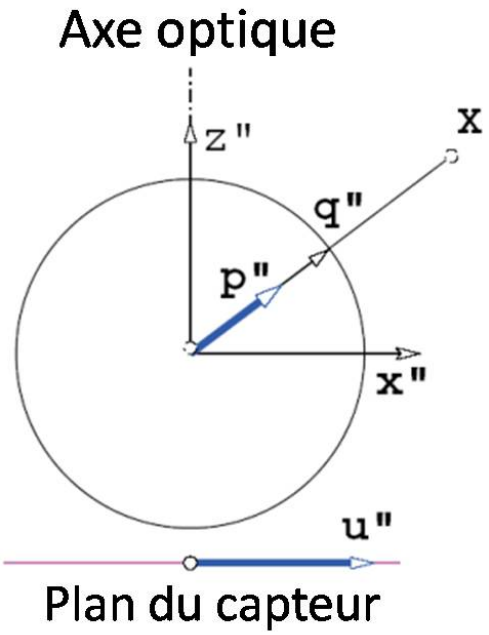


FIGURE 2.18 – La projection d'un point X du repère monde en un point u'' sur le plan du capteur [Micušık, 2004]

que le point $h(u'')$ est une projection orthographique du point p'' sur le plan du capteur, comme illustré sur la figure 2.19 (b).

Comme nous l'avons mentionné dans la section 2.1.2, le modèle unifié pour les caméras catadioptriques par Geyer et Daniilidis [18] démontre que les projections catadioptriques (parabolique, hyperbolique, elliptique) et perspectives sont isomorphes à deux projections successives. Une première projection centrale sur une sphère virtuelle, suivie d'une projection perspective sur le plan image. Dans le formalisme de Micusik [Micušık, 2004], la fonction g peut être considérée comme la fonction décrivant le profil du miroir, tandis que la fonction h peut être considérée comme la projection à travers la caméra perspective (par exemple pour la projection orthographique h = 1). La projection sur le plan du capteur

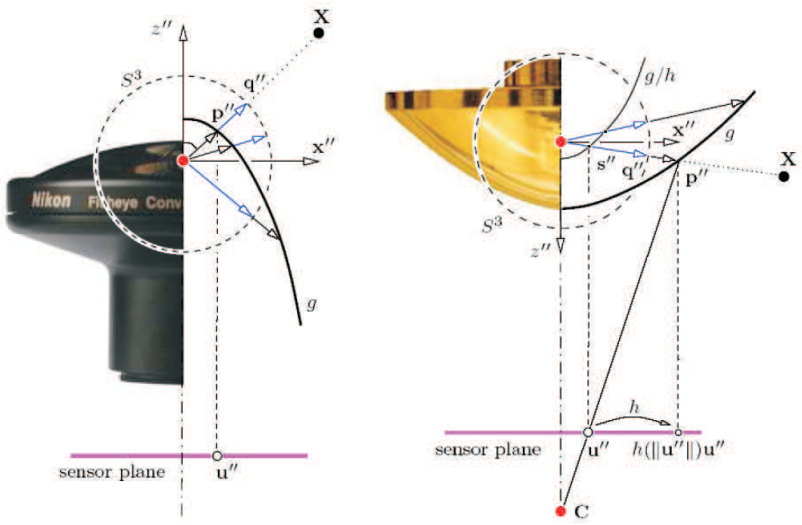


FIGURE 2.19 – Projection d'un point X du repère monde en un point u'' sur le plan du capteur : (a) capteur fiseyes et (b) capteur catadioptrique avec un miroir hyperbolique [Micušık, 2004]

peut être définie pour obtenir la projection perspective et la projection omnidirectionnelle comme suit :

$$\begin{aligned} \textit{Projection perspective} : & \begin{bmatrix} u'' \\ 1 \end{bmatrix}, \\ \textit{Projection omnidirectionnelle} : & \begin{bmatrix} h(\|u''\|)u'' \\ g(\|u''\|) \end{bmatrix}, \end{aligned} \tag{2.18}$$

avec

$$
\begin{aligned}
h(\|u^{''}\|) &= \frac{l(l+m)+\sqrt{\|u^{''}\|^2(1-l^2)+(l+m)^2}}{\|u^{''}\|^2+(l+m)^2} \\
g(\|u^{''}\|) &= \frac{l\|u^{''}\|^2+(l+m)+\sqrt{\|u^{''}\|^2(1-l^2)+(l+m)^2}}{\|u^{''}\|^2+(l+m)^2}
\end{aligned}
\tag{2.19}
$$

où les constantes l et m, dépendent du type de miroir (c'est-à-dire parabolique, hyperbolique, elliptique).

2.2.4 Conclusion

Dans ce chapitre nous avons vu les différentes contraintes auxquelles doit répondre un capteur omnidirectionnel composé d'un miroir et d'une caméra, pour être à point de vue unique. Les capteurs catadioptriques pouvant être utilisés, sont les capteurs composés d'un miroir parabolique avec une caméra orthographique, un miroir hyperbolique, plan, conique ou elliptique avec une caméra perspective. Cependant, comme nous l'avons vu précédemment et comme précisé dans [Pajdla et al., 2001], seuls les miroirs hyperboliques et paraboliques sont utilisés pour concevoir des caméras catadioptriques à point central unique. En effet, le miroir plan ne permet pas d'augmenter le champ de vision du capteur. Les miroirs coniques et sphériques sont des solutions dégénérées pour un système catadioptrique central car leur contrainte, ne permet pas de mettre en place un système réel (à point de vue unique). Le miroir elliptique n'est pas utilisé en pratique car, son champ de vision est inférieur à une demi-sphère de l'espace. On peut aussi ajouter les caméras fisheyes qui peuvent être modélisées par le modèle unifié en considérant $\xi > 1$. De plus, nous avons vu dans ce chapitre deux modèles possibles pour modéliser la formation des images omnidirectionnelles, le modèle unifié et le modèle polynomial mis en place par Scaramuzza. Dans la suite de cette thèse, lorsqu'on fera référence au modèle de formation des images, on considérera le modèle unifié basé sur une sphère unitaire virtuelle. Dans la

suite de nos travaux nous utiliserons une camera omnidirectionnelles composés d'un miroir hyperbolique et d'une camera perspective.

Chapitre 3

Détection 2D de personnes

Sommaire

3.1 État de l'art sur la détection de personnes dans les images

La détection de personnes dans les images statiques et les séquences vidéo est une tâche essentielle dans de nombreuses applications de vision par ordinateur, telles que l'interaction homme-robot, la navigation de robots en présence de personnes, la détection de piétons pour les systèmes d'aide à la conduite, ou encore la vidéo surveillance. Ce problème, est l'un des plus difficiles en vision par ordinateur en raison du grand nombre de situations possibles, comme par exemple les variations d'apparence et de pose des personnes ou encore l'occlusion partielle.

En raison de son importance et de ses nombreux défis, la détection de personnes est un domaine de recherche très actif. Dans cette section, différentes approches de détection de personnes sont présentées. Ces méthodes utilisent principalement des approches basées sur l'apprentissage d'un classificateur pour la création d'un modèle de classe prédéfinie, utilisé par la suite dans la phase de détection. Les méthodes de détection de personnes utilisent dans la majorité des cas des caméras perspectives. Dans les méthodes basées sur la classification, la détection est effectuée en deux étapes distinctes : L'apprentissage et la détection comme on peut le voir dans la (figure 3.1). L'étape d'apprentissage se concentre sur l'extraction de caractéristiques discriminantes de l'image. Plusieurs méthodes d'apprentissage peuvent être utilisées pour créer un modèle capable de discriminer la classe

"personne". Le modèle obtenu est ensuite utilisé pour la prise de décision lors de la phase de détection.

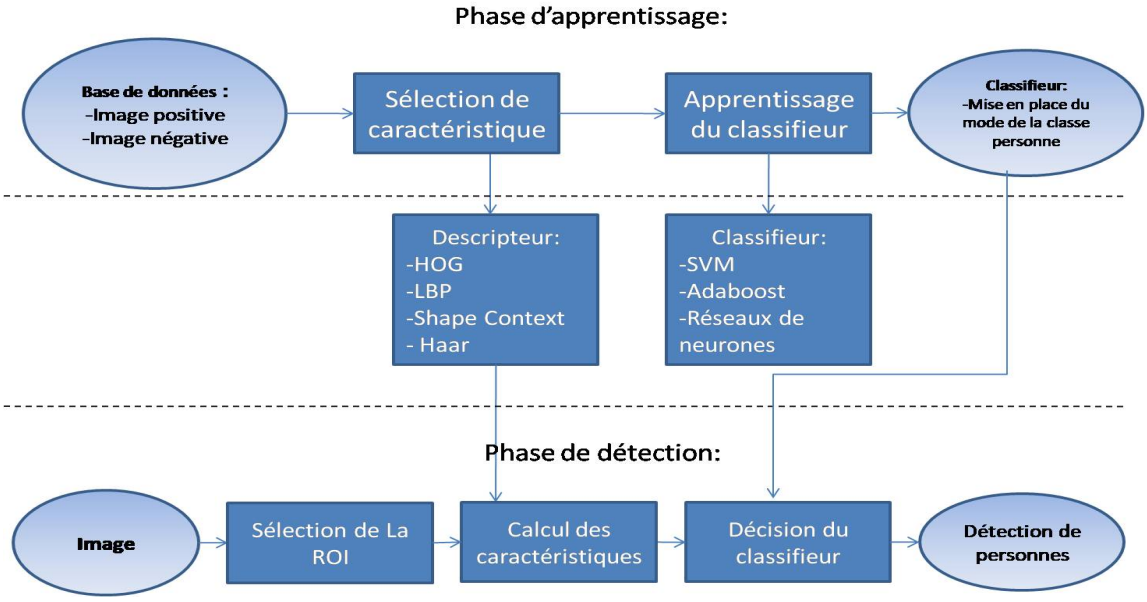


FIGURE 3.1 – Fonctionnement des méthodes basées descripteurs-classifieurs

3.1.1 Les descripteurs

Beaucoup de travaux ont porté sur la détection de personnes. Ils ont comme phase préalable, l'extraction de caractéristiques pertinentes afin de faire la distinction entre différentes classes. Or une image pour un ordinateur est représentée comme une matrice de pixels. L'utilisation de chaque pixel indépendamment, ne permet pas l'extraction d'informations pertinentes sur le contenu de l'image. De nombreuses approches d'extraction de caractéristiques visuelles utilisent des groupes de pixels. Ces caractéristiques peuvent être calculées à partir d'informations de bas niveau tels que le contour, la texture ou le mouvement quand il y a un mouvement dans la scène. Dans cette section, plusieurs ca-

ractéristiques sont revues en fonction des aspects de la forme humaine qu'elles décrivent : forme, texture et mouvement. Ces caractéristiques permettent d'extraire des informations significatives utilisées pour la détection de personnes.

3.1.1.1 Edge Orientation Histograms (EOH)

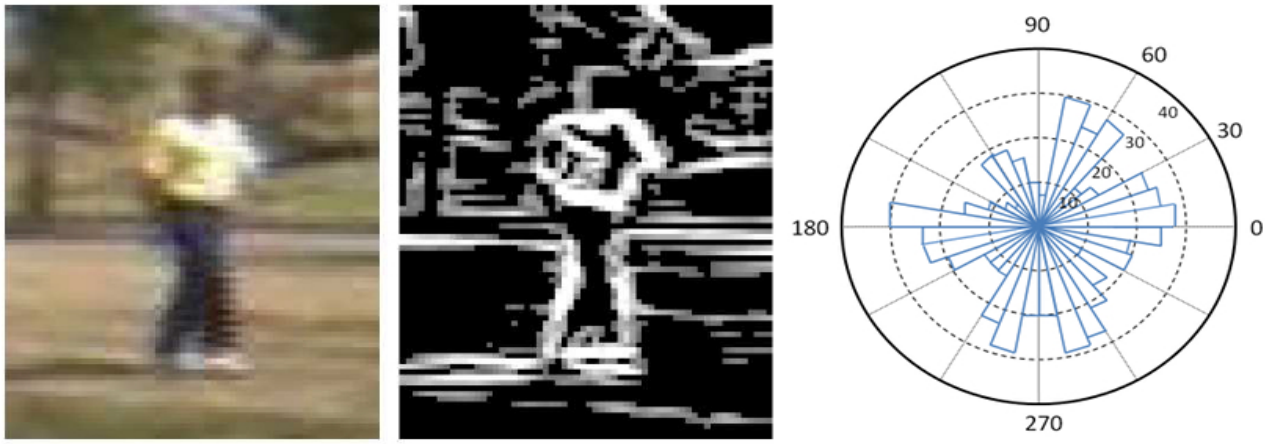


FIGURE 3.2 – Exemple des caractéristiques EOH

Les informations de silhouette et de contour sont des caractéristiques importantes pour discriminer une personne dans les images. Pour coder ces informations, les histogrammes orientés de bord (EOH), voir (figure 3.2) ont été proposés initialement pour la détection de visages par Levi et Weiss [Levi and Weiss, 2004]. Ces fonctionnalités permettent de conserver une invariance face aux changements globaux (sur toute l'image) de luminosité, mais aussi de décrire des propriétés géométriques difficiles à capturer avec d'autres descripteurs. Plus tard, les EOH ont été utilisés pour la détection des personnes. Dans [Gerónimo et al., 2007], une combinaison de caractéristiques Haar-like et d'histogrammes orientés de bord est utilisée comme caractéristique discriminante lors de la classification.

3.1.1.2 Les ondelettes de Haar

Les ondelettes de Haar ont été introduites pour la première fois dans le contexte de la détection d'objets à la fin des années 90 par Papageorgiou [Papageorgiou et al., 1998]. Viola et Jones [Viola and Jones, 2001] ont adapté l'idée d'utiliser des ondelettes Haar et ont développé les caractéristiques dites Haar. Ils ont introduit la notion d'image intégrale afin de calculer ces caractéristiques de manière rapide. Les caractéristiques de type Haar codent les relations entre les intensités moyennes des régions voisines selon des orientations différentes capturant des bords ou des changements de texture. Cela les rend cabables de détecter les similitudes structurelles entre différentes instances d'une classe. La figure (figure 3.3) montre les trois types de caractéristiques 2-D "Haar-like" utilisées par [Oren et al., 1997]. Ces caractéristiques captent le changement d'intensité locale selon les directions horizontales, verticales et diagonales. Lorsque ce détecteur est appliqué aux images, la valeur d'une caractéristique de deux rectangles est la différence entre la somme des pixels situés dans la zone non ombrée avec la somme des pixels situés dans la zone ombrée. Une caractéristique de quatre rectangles calcule la différence entre les paires diagonales de rectangles. Leinhart et al [Lienhart and Maydt, 2002] ont introduit un ensemble de fonctionnalités Haar étendues en ajoutant des fonctions rectangulaires orientées vers le haut, permettant aux prototypes d'être mis à l'échelle indépendamment dans les axes vertical et horizontal. La figure (figure 3.3) montre le modèle complet des caractéristiques de Haar utilisé par [Lienhart and Maydt, 2002].

3.1.1.3 Les motifs binaires locaux

Les motifs binaires locaux (Local Binary Pattern LBP) sont une caractéristique de codage de texture. C'est un cas particulier du modèle de texture proposé dans [Wang and He, 1990]

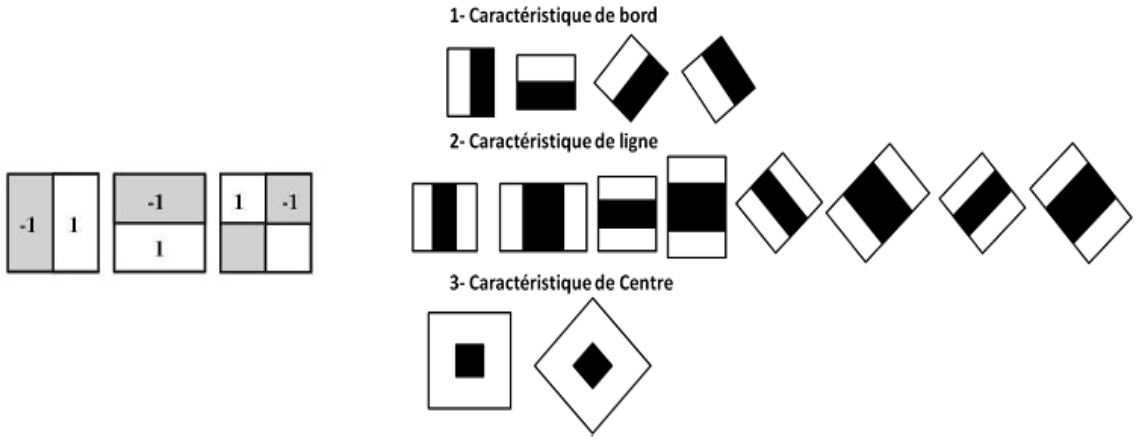


FIGURE 3.3 – Les ondelettes de Haar

et [He and Wang, 1990]. Il a été décrit pour la première fois dans [Ojala et al., 1996]. La version originale de la caractéristique des motifs binaires locaux pour chaque pixel est basée sur un bloc de 3x3 pixels d'une image. Les pixels de ce bloc sont définis par la valeur du pixel central, multipliée par des puissances de deux, puis additionnées pour obtenir une étiquette pour le pixel central. Comme le quartier se compose de 8 pixels, l'étiquette d'une valeur est comprise entre 0 et 256 en fonction des valeurs de gris du centre et des pixels dans le voisinage après la pondération comme dans la figure (figure 3.4). Une fonctionnalité LBP plus générique est proposée dans [Ojala et al., 2002]. Elle permet une meilleure extraction d'informations à partir du voisinage circulaire autour du pixel central, selon deux paramètres qui sont le rayon du cercle de voisinage "R" et le nombre de points de quartier considérés "P"(voir figure3.5).

De nombreuses approches de détection de personnes sont basées sur l'utilisation des descripteurs LBP uniquement, améliorant les performances de classification. Comme dans [Mu et al., 2008], où deux nouvelles variantes de la caractéristique LBP, sont proposées et utilisées pour la détection de personnes. Ces nouvelles variantes sont le LBP "sémantic-

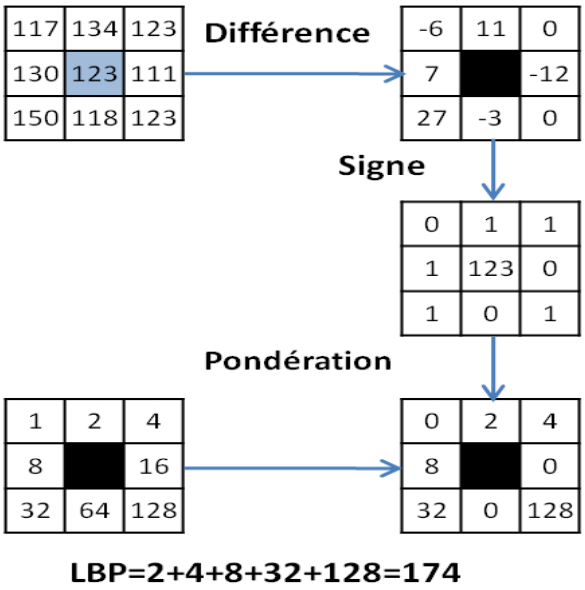


FIGURE 3.4 – Fonctionnement du descripteur LBP

LPB" (S-LBP) et le "Fourier-LBP" (F-LBP). Pour calculer ces descripteurs, il faut dans un premier temps effectuer une binarisation des pixels voisins dans un espace couleur tel que CIE-LAB. Ensuite pour le S-LBP, plusieurs ≪1≫ bits continus sur le cercle d'échantillonnage comme le montre la figure (figure 3.5) formant un arc qui va être représenté par sa direction principale et sa longueur d'arc. Les non uniformes (avec plus d'un arc) sont abandonnés. Un descripteur d'histogramme bidimensionnel (angle de l'arc par rapport à la longueur de l'arc) peut ainsi être obtenu pour une région d'image donnée.

Dans le cas du F-LBP, les caractéristiques sont conçues selon une idée similaire au descripteur de contour de Fourier [González et al., 2001]. Tout d'abord, les distances de couleur entre le pixel central et les pixels voisins sont regroupés dans un vecteur de caractéristiques, ensuite, ce vecteur de caractéristiques brut est transformé dans le domaine fréquentiel. Le coefficient de faible fréquence est conservé et utilisé pour la représentation

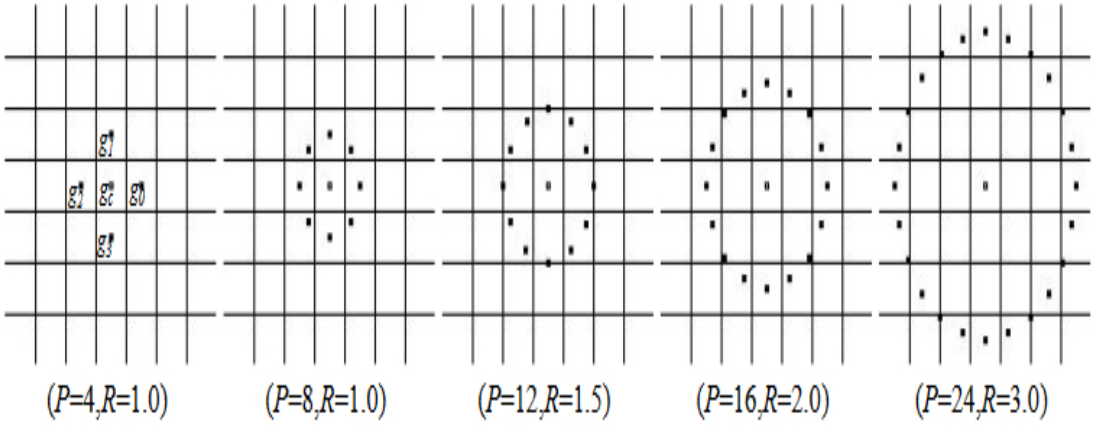


FIGURE 3.5 – Descripteur LBP générique

du F-LBP, puisqu'il capte les structures locales saillantes autour du pixel central. D'autres travaux utilisent les caracéristiques LBP combinées avec un autre descripteur comme dans [Zhou et al., 2012], où les caractéristiques HOG et LBP standards sont extraites des régions d'intérêt (ROI) du corps humain et sont combinées pour discriminer les personnes. La phase d'apprentissage se fait avec un classificateur SVM linéaire simple. Cette combinaison des deux descripteurs offre une meilleure performance de détection par rapport à l'utilisation de chaque descripteur indépendamment.

3.1.1.4 Contexte de forme

Les contextes de forme ont été introduits par [Belongie and Malik, 2000] pour des tâches de détection d'objets ou de reconnaissance de caractères. L'approche consiste à choisir n points sur les contours de forme de l'objet à classer. Le contour est extrait à l'aide d'un détecteur de contour, ensuite, chaque point du bord est décrit selon un diagramme de block "log-polaire" comme dans la figure (figure 3.6), l'algorithme a été testé sur la base

de données MNIST, et donne de bons résultats. Ces descripteurs sont très bien adaptés à des fins de correspondance et ont également été utilisés pour la détection de personnes dans [Leibe et al., 2005] ou pour la reconnaissance de caractères [Belongie et al., 2002] [Mori et al., 2005].

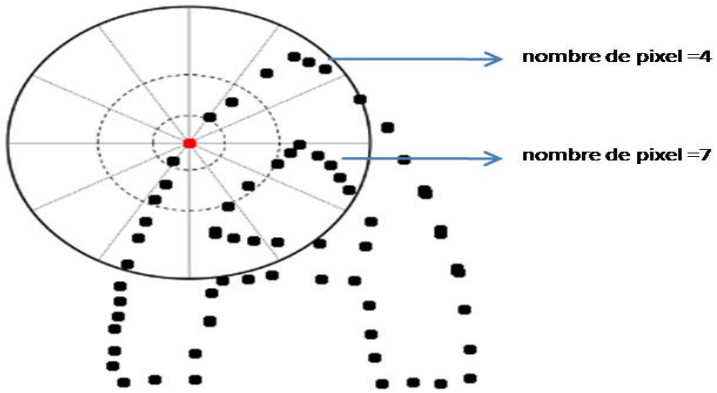


FIGURE 3.6 – Représentation de la distribution des pixels du contour dans chaque bloc

3.1.1.5 Histogramme Orienté du Gradient

Un autre descripteur extrêmement répondu basé sur la silhouette, est l'Histogramme Orienté du Gradient (HOG), proposé par Dalal et Triggs dans [Dalal and Triggs, 2005] pour la détection des personnes. L'extraction des caractéristiques est plus complexe que dans les Histogrammes d'orientation du bord, améliorant les performances discriminatoires du descripteur tout en assurant un certain degré d'invariance. Comme décrit dans [Dalal and Triggs, 2005], le calcul du descripteur HOG se fait en cinq étapes :

1. Une normalisation globale de l'image, en utilisant une compression gamma, est effectuée pour réduire l'influence que peut avoir une variation d'éclairage.

2. Calcul du gradient de l'image grâce à différents opérateurs (Roberts, Prewitt ou Sobel).

3. La fenêtre d'image est dans un premier temps divisée en petites régions spatiales, appelées «cellules» qui regroupent plusieurs pixels. On génère un histogramme d'orientation du contour en accumulant toutes les orientations de chaque pixel formant la cellule. Chaque pixel du contour contribue avec une valeur proportionnelle à la valeur de son orientation.

4. Une étape de normalisation s'effectue en accumulant une mesure de l'intensité de l'histogramme local sur des groupes de cellules appelées «blocs». Chaque cellule est normalisée par rapport au bloc auquel elle appartient. Les blocs se chevauchent et donc une même cellule peut contribuer à la formation de plusieurs blocs.

5. Le descripteur HOG de la fenêtre de détection est obtenu en concaténant tous les descripteurs HOG de tous les blocs. Ce vecteur permet de caractériser la forme de l'objet. L'extraction des caractéristiques HOG est illustrée dans la figure (figure 3.7) tirée de [Dalal and Triggs, 2005]. Quatre variantes du descripteur HOG ont été présentées par les auteurs. La différence entre eux, réside dans la forme des cellules considérées. Ces quatre variantes sont : HOG Rectangulaire (R-HOG), qui utilise une fenêtre rectangulaire, Circulaire HOG (C-HOG) où les cellules sont définies en grilles de forme log-polaire. Bar HOG où les descripteurs sont calculés de manière similaire au R-HOG, mais en utilisant des "bar filtres" que l'on peut calculer comme dans [Freeman et al., 1991] et des HOG de "Center-Surround" qui permettent de calculer les caractéristiques HOG de manière rapide, en utilisant la notion d' "Histogramme Intégral" (inspiré des "images intégrales" proposé par viola et al.) introduite par Porikli [Porikli, 2005]. Elle permet de calculer efficacement les histogrammes sur des régions d'image rectangulaires arbitraires. Dans [Zhu et al., 2006], zhu à utilisé cette notion pour mettre en place un descripteur HOG extrêmement rapide il peut traiter une image avec 800 fenêtres de détection en 26ms mais en contrepartie sa précision

est plus faible. Par la suite de nombreuses autres approches de détection de personnes, en utilisant des descripteurs HOG, ont été proposées. On peut citer [Bertozzi et al., 2007], [Zhu et al., 2006],[Abd-Almageed et al., 2007], [Alonso et al., 2007],[Shet et al., 2007], [Corvee and Bremond, 2010], [Felzenszwalb et al., 2010a], [Zaklouta and Stanciulescu, 2012] [Yang and Ramanan, 2013]. Elles diffèrent principalement dans la façon dont le descripteur HOG est utilisé (décrivant l'image entière ou des parties du corps indépendamment) ou par la méthode de classification utilisée (SVM, boosting, etc.).

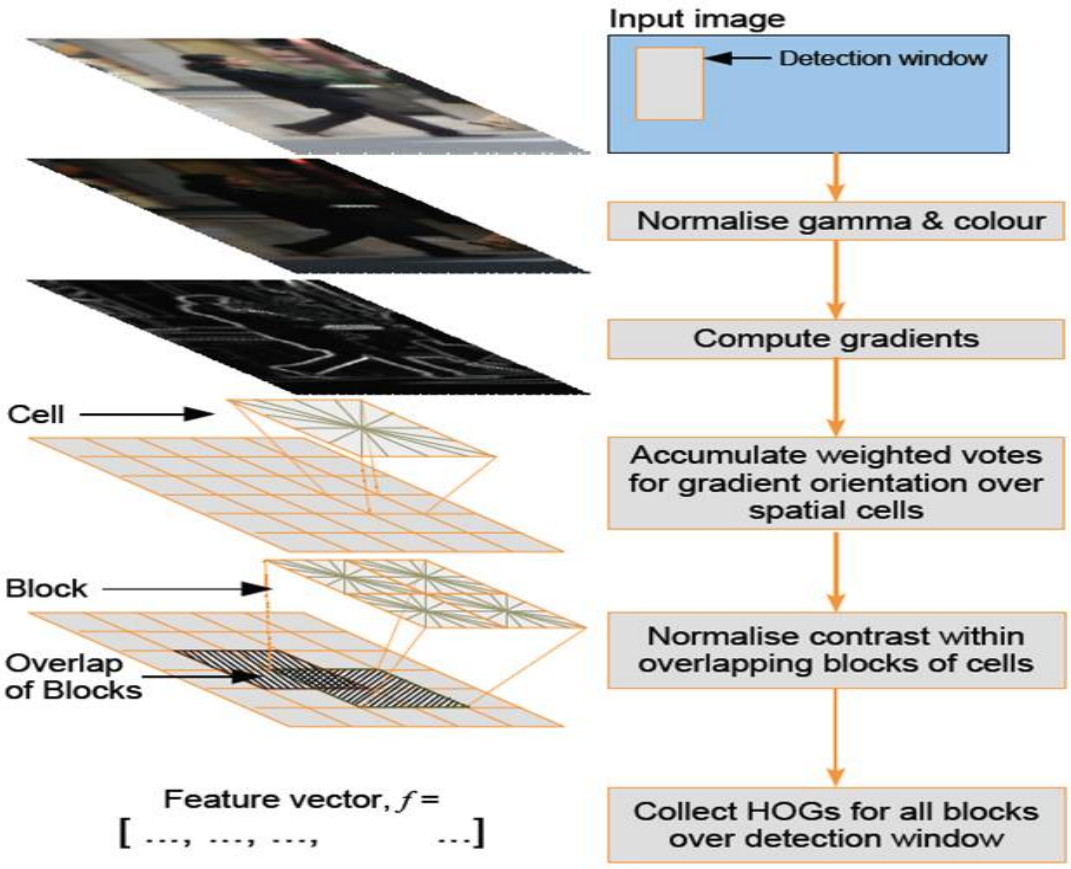


FIGURE 3.7 – Histogramme orienté du gradient pour la détection de personnes [dallal 2005]

3.1.2 Les classifieurs

La classification est la dernière étape de la détection de personnes. Au cours de la classification, une région candidate est évaluée et une décision est prise si cette région contient une personne ou non. Dans l'état de l'art la majorité des algorithmes de détection de personnes utilisent des approches basées sur des classificateurs. La plupart de ces approches utilisent des variantes de l'apprentissage par Boosting et SVMs. Mais d'autres techniques d'appariement de silhouettes peuvent être utilisées comme celles basées sur la distance de Chanfrein. Ces deux méthodes de classification sont présentées dans les paragraphes suivants.

3.1.2.1 Machine à vecteurs de support SVM

La fonction du classifieur est de donner une décision sur l'appartenance du candidat à la classe recherchée. Elle repose sur une base de données d'apprentissage. En prenant en entrée les caractéristiques des exemples contenant un individu de la classe (ici une personne) et des exemples ne contenant pas d'individus de la classe, le classifieur doit déterminer de qui les caractéristiques de l'image candidate sont les plus proches. Dans la plupart des cas, cette étape est la dernière du processus puisqu'une fois reconnues par le classifieur, il suffit d'afficher les fenêtres de détection.

Cette méthode élaborée par Vapnik et al [Cortes and Vapnik, 1995], vise à déterminer un hyperplan séparateur entre les espaces des deux classes (figure 3.8). L'idée est de maximiser la marge, c'est-à-dire la distance entre les frontières de séparation des échantillons les plus proches. Pour cela, l'algorithme transforme l'espace de représentation des données d'entrée en un espace de plus grande dimension, dans lequel il est probable qu'il existe une droite séparatrice linéaire. Du fait de sa très grande efficacité, cette méthode est très

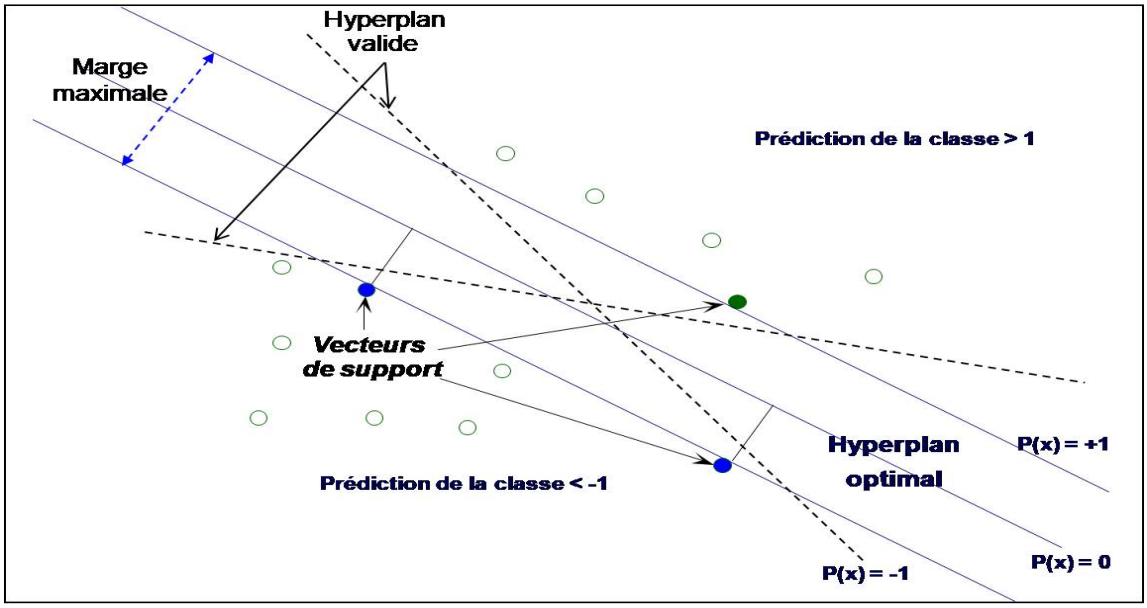


FIGURE 3.8 – Machine à vecteurs de support (SVM)

couramment utilisée notamment dans le domaine de la détection de personnes [Papageorgiou and Poggio, 2000], [Kang et al., 2002], [Jeong et al., 2004], [Sidenbladh, 2004], [Dai et al., 2005], [Felzenszwalb et al., 2010b] .

3.1.2.2 AdaBoost

AdaBoost (ou adaptive boosting) [Freund and Schapire, 1995] est un algorithme d'apprentissage utilisé dans [Viola and Jones, 2002], [Mikolajczyk et al., 2004], [Laptev, 2006]. Il peut être associé avec un autre algorithme d'apprentissage pour en améliorer les performances [Shashua et al., 2004] [Zhu et al., 2009]. Le principe consiste à combiner des classifieurs, par itérations successives. La connaissance d'un classifieur faible est ajoutée au classifieur final. Le classifieur ajouté est pondéré par la qualité de sa classification : plus il permet de bien classer, plus il sera important. Les exemples mal classés sont boostés pour

qu'ils aient davantage d'importance vis à vis de l'apprenant faible au prochain tour, afin de pallier à ce manque.

3.1.2.3 Réseaux de neurones

Un réseau de neurones est un algorithme d'apprentissage basé sur des concepts inspirés du fonctionnement d'un cerveau humain. Un réseau de neurones peut être considéré comme un graphe dirigé reliant un certain nombre de neurones par des liens pondérés, le fonctionnement des neurones est simulé par des fonctions d'activation. Les réseaux de neurones peuvent être classés en trois catégories basées sur l'architecture adoptée [Jain et al., 2000] : les réseaux feed forward (FFs) ou Perceptrons multicouches (MLP), les réseaux de fonctions à base radiale (RBFs) et les cartes auto-organisatrices (SOMs). Dans les réseaux MLP, une combinaison linéaire des entrées est calculée en utilisant une fonction qui renvoie le produit scalaire entre les entrées et les poids synaptiques correspondants. Par contre les RBFs emploient des fonctions de combinaison qui renvoient les distances euclidiennes entre les entrées et les centres de la couche cachée. Les cartes auto-organisatrices (SOMs) possèdent une structure différente où des noeuds (neurones) sont régulièrement placés dans une grille hexagonale ou rectangulaire. Les SOMs sont basés sur le principe de transformation des données d'entrée de grandes dimensions à des données de dimensions inférieures [Vesanto and Alhoniemi, 2000]. L'apprentissage d'un réseau de neurones implique l'ajustement de poids basé sur un processus itératif qui optimise une fonction d'erreur particulière. Les méthodes de la descente de gradient par exemple sont des méthodes d'apprentissage supervisées qui ont été beaucoup utilisées grâce aux résultats raisonnables qu'ils peuvent générer en un temps minimal. Cependant, elles peuvent converger vers des solutions locales dans la plupart des cas et plus particulièrement lorsque la tâche est difficile, telle que la reconnaissance de caractères manuscrits. Les réseaux de neurones, avec leur capacité de

généralisation, peuvent reconnaître et détecter des formes imprévues qui sont difficiles à détecter par d'autres méthodes de classification. Un réseau de neurones entraîné peut être considéré comme un "expert" capable de fournir des réponses dans de nouvelles circonstances (une situation inédite).

Comme nous avons vu dans cette section, différents types de classifieurs permettent de discriminer les différentes classes. Lors de la phase d'apprentissage, un classifieur doit disposer d'une base de données annotée comme celle de l'INRIA dans le cas d'images perspectives, lui permettant de mettre en place le modèle de classe. Dans la section suivante, nous verrons la mise en place d'une base de données dans le cas des images omnidirectionnelles que nous utiliserons pour notre détection.

Nous avons présenté les différentes approches de détection de personnes mises en place en vision perspective et omnidirectionnelle. Dans la suite de ce chapitre, nous présenterons notre système de détection de personnes qui devra par la suite initialiser un algorithme de suivi 3D, tout en réduisant la région d'intérêt. Notre détecteur de personnes dans les images omnidirectionnelles, n'est qu'une partie dans un traitement plus complexe qui permet l'estimation de pose 3D. Cela entraîne des contraintes supplémentaires sur le temps de calcul afin de permettre la mise en place d'un système réactif.

3.2 Génération de la base de données omnidirectionnelles

La base de données d'images a un impact central sur l'amélioration de l'efficacité des algorithmes de détection. Plusieurs bases de données pour la détection des personnes sont proposées dans la littérature comme celle de l'INRIA. Cependant, elles sont toutes adaptées aux images conventionnelles et jusqu'à présent, à notre connaissance, aucune base de données d'images omnidirectionnelles pour la détection de personnes n'a été réalisée.

Nous proposons une technique pour créer une base de données adaptée pour le cas des images omnidirectionnelles, à partir d'une base de données perspective. Soit I_c une image perspective de la base de données conventionnelle. Nous considérons cette image comme un objet plan observé par une caméra omnidirectionnelle, comme on peut le voir sur la figure (figure 3.9) . L'image omnidirectionnelle correspondante peut être utilisée pour générer la base de données adaptée. En effet, l'image omnidirectionnelle d'une personne est générée sous la contrainte de visibilité globale de la personne vue par la caméra omnidirectionnelle. Cette hypothèse peut être vérifiée pour une distance suffisante entre la caméra et la projection de l'image perspective dans un espace monde virtuel. La technique proposée est expliquée dans les étapes suivantes :

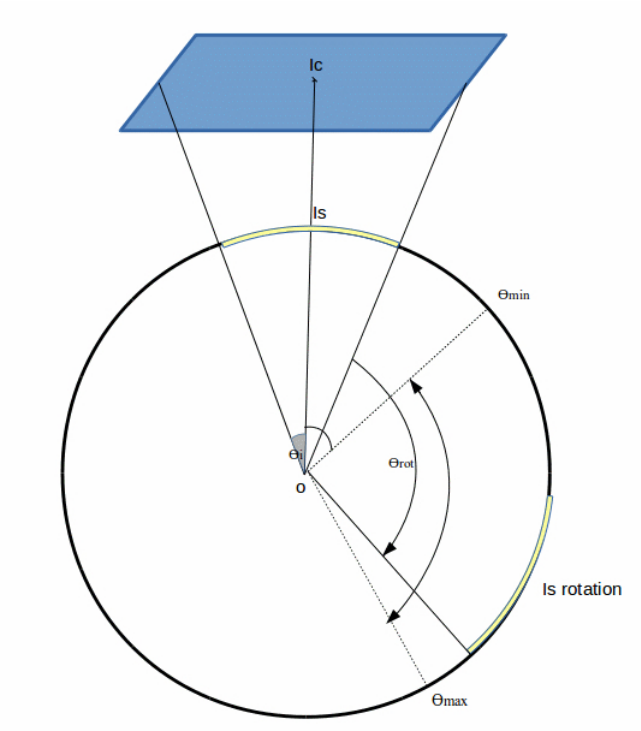


FIGURE 3.9 – Image perspective projetée sur la sphère

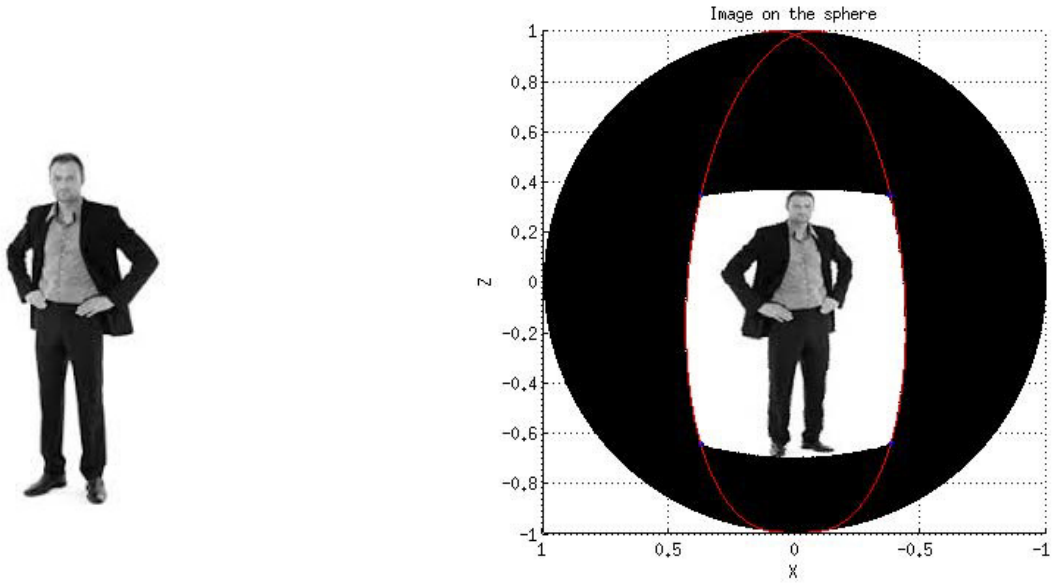
Tout d'abord, une caméra perspective virtuelle est définie, avec comme matrice des paramètres intrinsèques K_v déterminée. Nous utiliserons K_v pour passer de l'image I_c vers le plan normalisé I_n suivant le modèle sténopé classique. I_c est définie à une distance adéquate de la sphère unitaire et le vecteur normal au plan contenant I_c est colinéaire à l'axe passant par le pôle nord comme illustré sur la figure (figure 3.9). On pose un point P_n appartenant à l'image I_n et un point P_c appartenant à l'image I_c.

$$P_n = K_v^{-1} P_c \tag{3.1}$$

Ainsi nous passons de l'image pixélisée I_c à l'image normalisée I_n. Ensuite, l'image normalisée générée est projetée sur la sphère unitaire pour obtenir une image sphérique :

$$P_s = \frac{P_n}{\|P_n\|} = (x_s, y_s, z_s) \tag{3.2}$$

Nous notons que la base de données adaptée est générée et définie sur l'espace sphérique. L'espace sphérique présente l'avantage d'être invariant à la rotation et donc qu'une seule image sphérique (θ, φ) est générée pour chaque image perspective comme on peut le voir sur la figure (figure 3.10). Ce qui ne sera pas le cas lorsque l'on voudra générer des images omnidirectionnelles. Il faudra faire la distinction entre les rotations autour de l'axe φ et les rotations autour de l'axe θ. La rotation autour de l'axe φ dans l'espace sphérique se traduit dans l'image omnidirectionnelle par une rotation autour du centre et qui par conséquent ne modifie pas les distorsions créées dans l'image projetée. Par contre les rotations autour de l'axe θ dans la sphère vont se traduire dans l'image omnidirectionnelle par un mouvement du centre de l'image vers le cercle extérieur (ou l'inverse). Or dans l'image omnidirectionnelle, les distorsions ne sont pas les mêmes au centre de l'image omnidirectionnelle et sur les bords. Pour intégrer ce phénomène dans notre base de données d'images omnidirection-



(a) (b)

FIGURE 3.10 – Transformation d'une image perspective (a), en image sphérique (b).

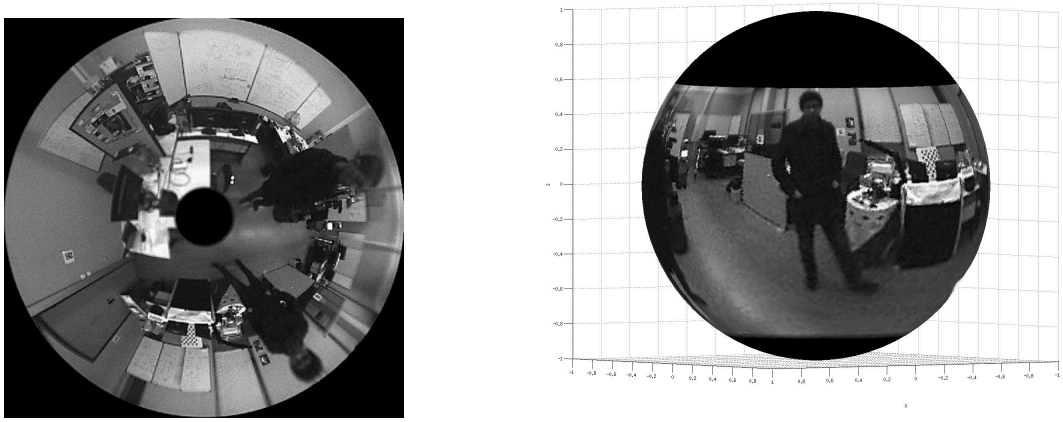
nelles, nous définissons θ_{min}, θ_{max} qui sont les limites de la zone de visibilité de notre image omnidirectionnelle comme dans la figure (figure 3.9). Ainsi, nous utiliserons les équations (3.1) et (3.2), pour obtenir notre image I_o qui représente l'image omnidirectionnelle comme suit :

$$I_o = K_{vo} * h \left(I_s * \begin{bmatrix} cos(\phi) & -sin(\phi) & 0 \\ sin(\phi) & cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & -sin(\theta) \\ 0 & sin(\theta) & cos(\theta) \end{bmatrix} \right) \tag{3.3}$$

Avec K_{vo} les paramètres intrinsèques de notre caméra omnidirectionnelles. Le nombre d'images omnidirectionnelles générées à partir d'une seule image perspective sera égal à θ_{rot} divisé par le pas que nous aurons choisi. Comme on peut le voir sur la figure (figure

FIGURE 3.11 – Images omnidirectionnelles obtenues à partir d'une image perspective pour différentes rotations θ.

3.11), plus l'image se rapproche du bord, plus les distorsions sont importantes (visible). Les images sphériques obtenus ne sont pas des images acquises par la caméra et cela nécessite donc une phase d'interpolation des pixels pour obtenir un échantillonnage régulier sur la sphère est nécessaire. En pratique, cette étape est effectuée par une interpolation linaire ou bilinéaire dans le plan image de la caméra qui n'est théoriquement pas correcte et engendre donc des artéfacts (figure 3.12).



(a) (b)

FIGURE 3.12 – Exemple d'interpolation permettant de visualiser une image sphérique (b) de l'image omnidirectionnelle (a).

3.3 Descripteur HOG dans les images omnidirectionnelles

3.3.1 Calcul du gradient dans les images catadioptriques

3.3.1.1 Variété Riemannienne

Dans [Puig et al., 2014] et [Bogdanova et al., 2007], l'opérateur différentiel dans une variété riemannienne est utilisé pour le calcul du gradient. Soit $\mathcal{M}$ une variété différentielle

(a) Omnidirectional Image (b) Calcul du gradient avec la métrique Riemannienne de l'image (a) (c) Visualisation de HOG

FIGURE 3.13 – Exemple de descripteur HOG dans les images omnidirectionnelles.

de la surface R_3 avec une métrique riemannienne induite g_{ij} qui encode les propriétés géométriques de la variété. x^i est le système de coordonnées dans $\mathcal{M}$.

Le gradient associé à une métrique riemannienne est défini comme suit :

$$\nabla f = \sum_{i=1}^{n} \sum_{j=1}^{n} g^{ij} \frac{\partial f}{\partial x^j} \frac{\partial}{\partial x^i} \tag{3.4}$$

Où g^{ij} est l'inverse de la métrique riemannienne g_{ij}.

Un point de la sphère unitaire S_2, peut être représenté selon les coordonnées cartésiennes ou polaires par : $(X, Y, Z)^T = (sin\theta sin\phi, sin\theta cos\phi, cos\theta)^T$

Le déplacement élémentaire euclidien est défini comme suit :

$$dl^2 = dx^2 + dy^2 + dz^2 = d\theta^2 (sin\theta)^2 d\phi \tag{3.5}$$

Un point sur la sphère est représenté par (θ, ϕ). Il devient après une projection stéréographique un point (R, ϕ) dans le plan image. L'angle θ dépend des paramètres de calibrage de notre caméra, tandis que ϕ reste le même.

Dans le cas général nous avons :

$$\theta = arctan\left(\frac{R\left(1+\xi+\sqrt{(1+\xi)^2 - R^2\left(\xi^2-1\right)}\right)}{1+\xi-R^2\xi+\sqrt{(1+\xi)^2 - R^2\left(\xi^2-1\right)}}\right) \tag{3.6}$$

Ainsi, la métrique devient :

$$dl^2 = \frac{\left(\xi+\xi^2+\sqrt{(1+\xi)^2 - R^2\left(\xi^2-1\right)}\right)^2\left(R^2d\phi^2+\frac{(1+\xi)dR^2}{1-R^2(\xi-1)+\xi}\right)}{\left(R^2+(1+\xi)^2\right)^2} \tag{3.7}$$

En utilisant les identités : $R = x^2+y^2$ et $\phi = tan^{-1}\left(\frac{y}{x}\right)$ nous avons :

$$dl^2 = \lambda\left(4\left(ydx-xdy\right)^2 - \frac{4\left(1+\xi\right)\left(xdx+ydy\right)^2}{\left(x^2+y^2\right)\left(\xi-1\right)-\xi-1}\right) \tag{3.8}$$

avec :

$$\lambda = \frac{\left(\xi+\xi^2+\sqrt{\left(1+\xi^2\right)^2 - \left(x^2+y^2\right)\left(\xi^2-1\right)}\right)^2}{4\left(x^2+y^2\right)\left(x^2+y^2+(1+\xi)^2\right)^2} \tag{3.9}$$

A partir des équations précédentes, on peut calculer g^{ij} :

$$g^{ij} = \gamma\begin{pmatrix}-x^2(\xi-1)+\xi+1 & xy(\xi-1)\\ xy(\xi-1) & -y^2(\xi-1)+\xi+1\end{pmatrix} \tag{3.10}$$

avec :

$$\gamma = \frac{\left(x^2+y^2+(1+\xi)^2\right)^2}{(1+\xi)\left(\xi+\xi^2+\sqrt{1-\left(x^2+y^2\right)\left(\xi^2-1\right)+2\xi+\xi^2}\right)^2} \tag{3.11}$$

Le gradient basé sur la métrique riemannienne est rapide à calculer. Il peut être vu comme une simple pondération de la fonction du gradient conventionnel, calculé dans l'image omnidirectionnelle avec le paramétré ξ qui tient compte de la forme du miroir. La figure (figure 3.13(b)), montre le calcul du gradient dans une image omnidirectionnelle réelle. Son descripteur HOG est présenté par la (figure 3.13(c)).

3.3.1.2 Gradient Sphérique

Le gradient sphérique peut être obtenu en prenant en compte la métrique dans la sphère unitaire $r = 1$. En partant de $g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & sin\theta \end{pmatrix}$ on a $g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{sin\theta} \end{pmatrix}$
Le gradient sphérique est ainsi défini par :

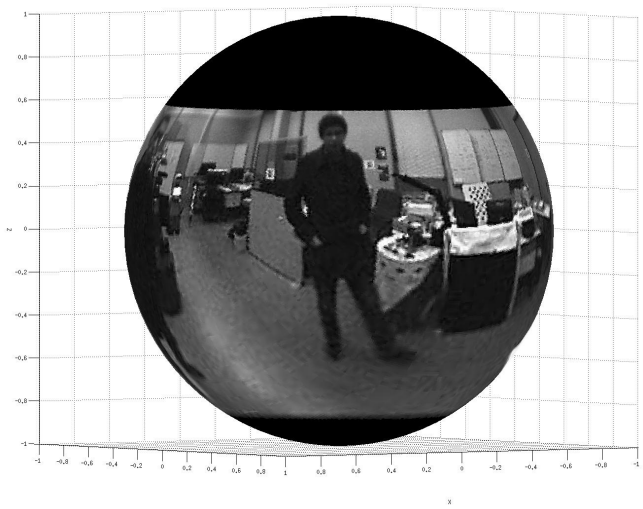
$$\nabla_{S^2} I_s(\theta, \phi) = \frac{\partial I_s(\theta, \phi)}{\partial \theta} e_\theta + \frac{1}{sin\theta} \frac{\partial I_s(\theta, \phi)}{\partial \phi} e_\phi \tag{3.12}$$

Où (θ, ϕ) de l'image sphérique (θ, ϕ) sont respectivement les angles de longitude et de colatitude et e_θ et e_ϕ sont des vecteurs unitaires.
En pratique, le gradient est calculé en utilisant les dérivées de premier ordre de l'image comme suit :

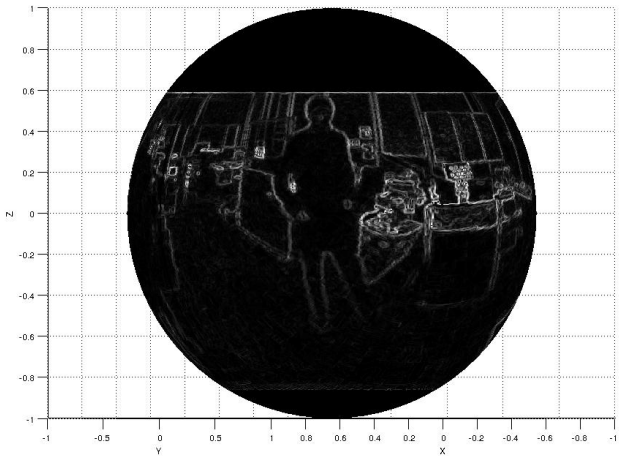
$$(\nabla_{S^2} I_s(\theta_i, \phi_j))^2 = |I_s(\theta_{i+1}, \varphi_i) - I_s(\theta_i, \varphi_{j+1})|^2 \tag{3.13}$$

$$+ \frac{1}{sin^2(\theta_j)} |I_s(\theta_{j+1}, \varphi_{k+1}) - I_s(\theta_j, \varphi_{k+1})|^2$$

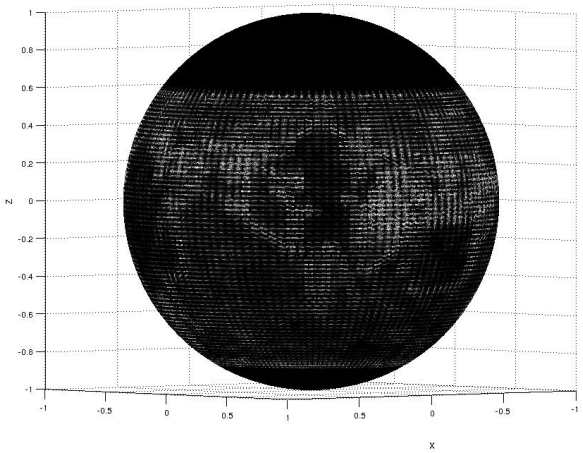
Une fois que le gradient est calculé sur la sphère, nous pouvons générer le descripteur HOG dans l'espace sphérique. La (figure 3.14(b)) montre le calcul du gradient sphérique d'une image omnidirectionnelle réelle. Son descripteur HOG est représenté dans la (figure 3.14(c))



(a) Image omnidirectionnelle représentée sur la sphere unitaire



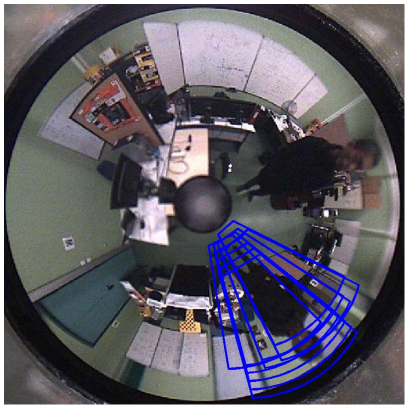
(b) Gradient sphérique de l'image (a)



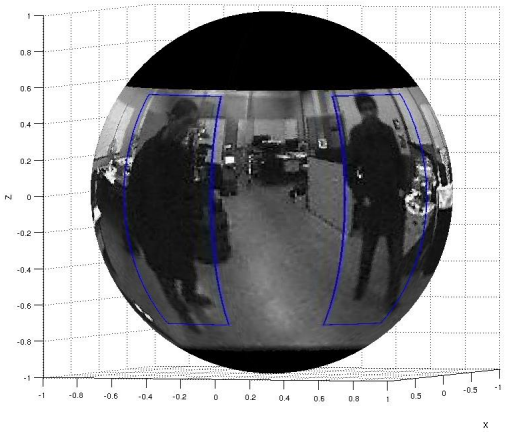
(c) Visualisation du HOG sphérique calculée dans l'image (θ, φ) (b)

3.4 Résultats

(a) Détection dans l'images dépliée



(b) Détection dans l'image omnidirectionnelle



(c) Détection dans l'image sphérique

FIGURE 3.15 – Un exemple de détection avec les trois méthodes

Afin de valider notre approche, nous utilisons la base de données perspective de l'INRIA de prés de 1500 images positives (contenant une personne). Au cours de nos expériences, nous comparons trois méthodes. La première méthode est basée sur l'utilisation d'un HOG

classique. Pour cette méthode, nous utilisons la base de données INRIA afin de former un SVM linéaire. Dans la deuxième méthode, HOG est utilisé avec un calcul adapté du gradient pour le cas d'une image omnidirectionnelle, grâce à la métrique de Riemann. Lors de la phase d'apprentissage de cette méthode, la base de données générée avec la méthode GDO présentée dans la section précédente a été utilisée. Chaque image en perspective génère 11 images omnidirectionnelles, pour l'étape de formation. Nous avons généré au total 16500 images d'apprentissage positives. Le paramètre d'étalonnage de notre caméra omnidirectionnelle virtuelle ainsi que la taille de l'échantillonnage de la sphère unitaire ont été conçus de telle sorte qu'une personne à une certaine distance par rapport aux capteurs soit vue entièrement dans l'image. Les images négatives sont générées à l'aide d'images omnidirectionnelles négatives. Car une seule image omnidirectionnelle sans la présence d'une "personne" permet d'obtenir beaucoup d'images négatives (selon le pas de rotation de notre fenêtre). Pour la phase d'apprentissage de la troisième méthode basée sur le gradient sphérique, le traitement est effectué sur les images sphériques $I(\theta, \varphi)$. L'avantage de cette méthode est que, lors de la transformation des images perspectives, de la base de données INRIA, une image en perspective générera une seule image sphérique. Ceci est dû au fait que l'image sphérique est invariante à la rotation. Donc nous aurons 1500 images positives.

Pendant la phase d'apprentissage de SVM, nous avons divisé notre base de données créée en deux parties. La première nous permet de faire un apprentissage initial, la seconde nous aide à renforcer cet apprentissage en effectuant un second apprentissage "hard examples". Cette méthode consiste à tester SVM formé grâce à la première partie sur la deuxième partie de la base de données. Les images mal détectées dans cette seconde partie seront intégrées une nouvelle fois à l'apprentissage permettant d'améliorer les performances de notre classifieur. Pour la base de données de test, nous utilisons 130 images acquises

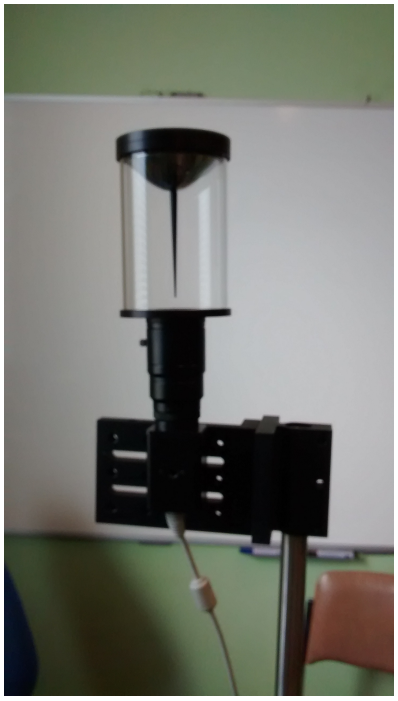
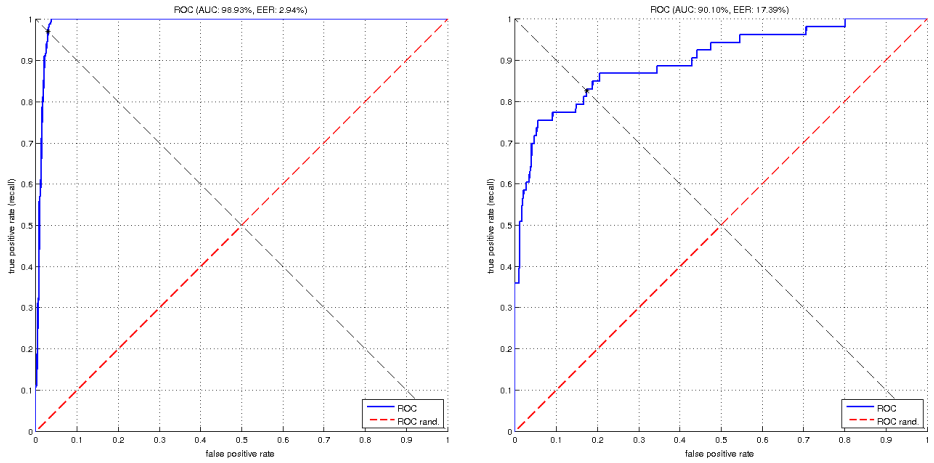


FIGURE 3.16 – La caméra catadioptrique avec un miroir hyperbolique de notre protocole expérimental.

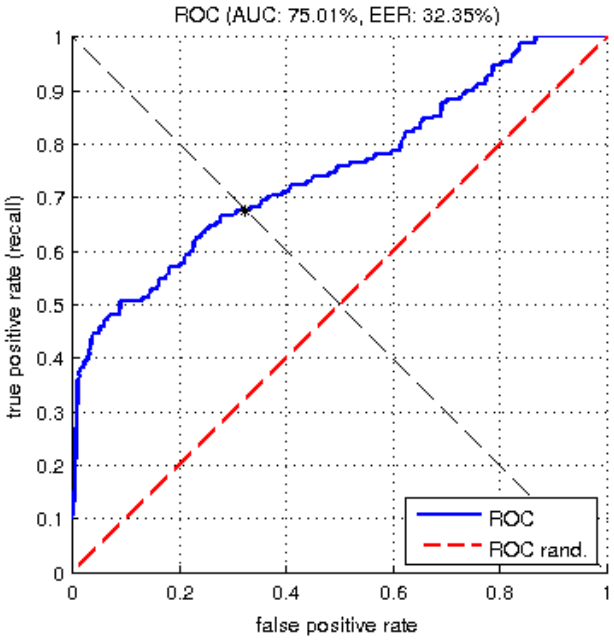
directement à l'aide d'une caméra omnidirectionnelle avec un miroir hyperbolique (figure 3.16). Toutes les images acquises ont été annotées manuellement pour être considérées comme notre vérité terrain. Pour quantifier les performances de notre détecteur, nous nous sommes appuyés sur les courbes ROC (Receiver Operating Characteristic). L'axe des abscisses représente le taux de vrai positif (recall) défini comme suit : $\left(\frac{True\ Pos}{True\ Pos+False\ Neg}\right)$ et l'axe des ordonnées le taux de faux positive : $\left(\frac{False\ Pos}{False\ Pos+True\ Neg}\right)$. Cela nous permet de comparer facilement nos 3 méthodes, en prenant en compte la surface sous la courbe (AUC Area Under the Curve).

Une fenêtre détectée est considérée comme un vrai positif si elle couvre au moins 50% de



(a) Resultats Méthode 3 : gradient sphérique

(b) Resultats Méthode 2 : gradient métrique riemannien



(c) Resultats Méthode 1 : HOG classique

FIGURE 3.17 – Résultats de la détection avec les trois méthodes.

la surface de notre annotation.

$$\frac{area\,(Detected\ Window \cap Annotated\ Box)}{area\,(Detected\ Window \cup Annotated\ Box)} > 0.5 \tag{3.14}$$

Dans un premier temps, nous avons comparé les trois méthodes (figure 3.17). La première chose que l'on constate, c'est que l'utilisation de HOG classique sur des images omnidirectionnelles dépliées, donne de médiocres performances. Les deux méthodes adaptées aux images catadioptriques permettent d'améliorer de manière significative les performances de notre détecteur. En particulier, la méthode basée sur la sphère unitaire, qui permet une amélioration significative des performances de détection : 7,53 % (représente l'augmentation de la surface sous la courbe) par rapport à la méthode basée sur le calcul du gradient avec la métrique riemannienne. Cela nous semble dû principalement au fait que

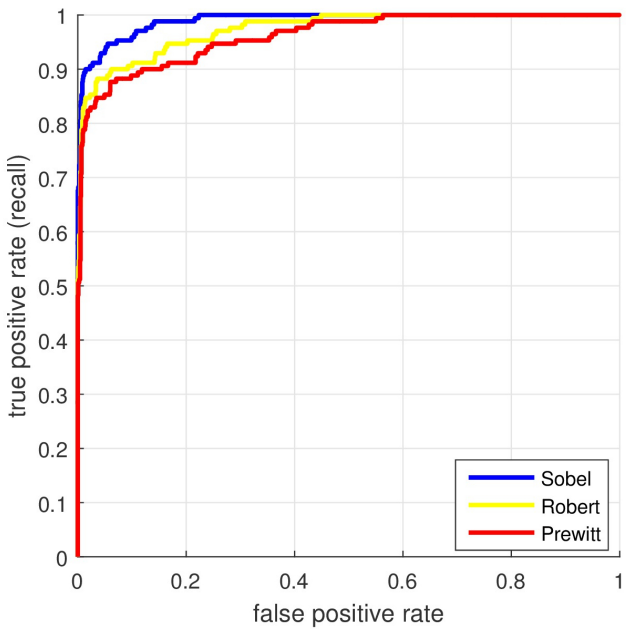


FIGURE 3.18 – Résultats selon le filtre utilisé pour le calcul du gradient

dans l'espace sphérique, l'image sphérique est invariante à la rotation et donc l'image de la personne sera la même qu'importe sa position autour du capteur. Les images de détection et d'apprentissage subissent les mêmes distorsions.

Ayant validé les performances du HOG sphérique, nous avons également testé les différentes possibilités pour le détecteur sphérique. Nous avons comparé les trois descripteurs sphériques selon le filtre sur lequel ils sont basés à savoir : Sobel, Robert et Prewitt. Comme on peut le voir sur la (figure 3.18), celui basé sur le filtre Sobel donne de meilleures performances. En effet, il améliore de 2 % la détection dans notre base de données.

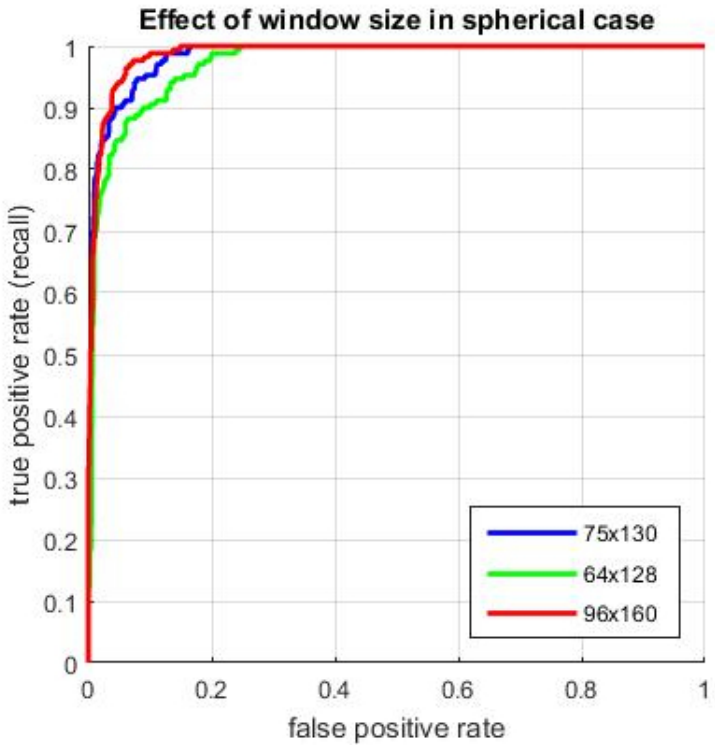


FIGURE 3.19 – Résultats selon la taille de la fenêtre dans le cas sphérique

Nous établissons, que le choix de la taille de la fenêtre peut être important dans la

configuration de notre descripteur. Nous avons constaté que les meilleures performances sont obtenues avec l'utilisation de fenêtre de taille 96x160, comme on le voit sur la figure (figure 3.19).

3.5 Conclusion

Dans ce chapitre, nous avons décrit un système efficace pour la détection de personne avec une caméra omnidirectionnelle, on utilisant le gradient sphérique combinée avec un SVM à noyau linéaire. Nous avons présenté notre approche qui permet une amélioration significative de la performance de l'algorithme HOG pour la détection de personnes dans des images omnidirectionnelles en utilisant la représentation sphérique. Ceci est possible grace à l'utilisation du modèle unifié. Nous avons étudié l'influence de divers paramètres descripteurs et avons conclu, que le filtre et la taille des fenêtres sont importants pour une bonne performance.. Les résultats expérimentaux présentés dans ce document confirment l'efficacité et la robustesse de l'approche proposée.

De plus lors de la mise en place de la partie expérimentale il a été nécessaire de créer une base de données omnidirectionnelles pour effectué les tests. Nous avons introduit une méthode permettant de convertir les images perspectives en images omnidirectionnelles pour transformer une base de données perspective en base de donnée omnidirectionnelle. Ainsi dans l'approche proposée au cours de l'étape d'apprentissage nous n'avons pas besoin de créer des images multiples pour une seule personne perspective en fonction de sa position dans l'image, puisque l'image sphérique est invariante à la rotation.

Chapitre 4

Suivi 3D de personnes

Sommaire

La recherche de l'estimation du mouvement 3D d'une personne est un domaine important de la vision par ordinateur, de par ses nombreuses applications possibles : interfaces homme-ordinateur, animation, interaction avec des environnements virtuels, jeux, etc. Capturer le mouvement humain 3D en temps réel, avec une seule caméra et sans marqueurs est une chose difficile. Cela est du aux ambiguïtés résultant du manque d'informations de profondeur, d'occultations partielles de parties du corps humain, du nombre élevé de degrés de liberté, de la variation dans les proportions du corps humain, ainsi que de la couleur des vêtements des différentes personnes présentes dans la scène. Pour ces raisons, le nombre de travaux traitant de l'estimation de pose de personnes continue d'augmenter. Dans ce chapitre, nous présenterons les différentes approches d'estimation de pose et du suivi de mouvements de personnes. Dans la première partie, un bref état de l'art du suivi 3D de personnes, colligé dans la littérature pour les images perspectives et omnidirectionnelles, sera présenté. Les méthodes d'estimation de pose 3D seront passées en revue dans un contexte général, mettant l'accent sur la partie suivi visuel, en particulier, sur l'utilisation du filtrage particulaire dans le suivi 3D. Ensuite, nous présenterons les différentes méthodes de modélisation du corps humain. Dans la dernière section nous exposerons notre contribution sur la mise en place de fonctions de vraisemblance, basées sur les distances

géodésiques dans l'espace de la sphère $\left(S^2\right)$, qui permettent de faire un suivi 3D dans des images omnidirectionnelles. Nous présenterons aussi les expériences effectuées sur une base de données créée pour le suivi 3D d'images omnidirectionnelles.

4.1 Approche 3D

Dans cette partie nous décrirons les approches proposées pour la reconnaissance de pose d'une personne dans l'espace 3D. L'approche générale consiste à trouver les paramètres d'un modèle 3D, de telle façon que la projection de ce dernier sur le plan de l'image soit recalée sur la silhouette de la personne détectée. Nous ferons la distinction par rapport aux travaux précédents en fonction du nombre de caméras requis pour les différentes approches.

4.1.1 Approche Multi-caméras

Pour améliorer la précision des mesures 3D et résoudre les ambiguïtés de l'occlusion, certaines approches proposent l'utilisation de plusieurs caméras dans le processus de reconnaissance de pose. La majorité des approches existantes peuvent être classées comme des approches basées sur des modèles. Elles consistent à retrouver les paramètres du modèle humain 3D, en effectuant la projection du modèle dans les différents plans image correspondant à la silhouette réelle. Dans [Delamarre et Faugeras, 2001], les auteurs proposent un modèle humain 3D constitué de cônes tronqués pour les articulations, de sphères pour la tête et de parallélépipèdes pour le tronc afin de s'adapter au mieux à la personne observée par trois caméras. Le modèle possède 22 degrés de liberté correspondant aux différentes articulations du corps humain. Leur algorithme calcule la "force" nécessaire pour faire correspondre le contour du modèle 3D projeté sur le plan de l'image avec le contour de

la personne détectée et cela pour chaqu'une des caméras. Ces forces sont ensuite additionnées et appliquées au modèle 3D. En outre, les auteurs supposent que la pose initiale est connue. Ainsi la posture à l'instant t-1, initialise la pose à l'instant t. Les travaux plus récents, utilisent généralement le filtre à particules (PF) qui est une technique stochastique pour l'estimation et le suivi 3D. Le filtrage particulaire, [Arulampalam et al., 2002], [Lee and Elgammal, 2010] est l'une des approches courantes pour le suivi du mouvement humain. Il utilise la pose courante et un modèle dynamique pour prédire la pose suivante. Le filtre à particules, utilise de multiples prédictions. Il propage des particules où chaque particule représente une position possible du modèle généré en utilisant le modèle dynamique. Ces particules sont ensuite comparées aux données observées dans l'image grâce aux fonctions de vraisemblance. La pose préalable est généralement assez diffusée mais la fonction de vraisemblance du modèle dynamique permet de déterminer les positions les plus probables. La principale difficulté de l'application du filtre à particules est la grande dimension de l'espace d'état dans l'estimation de la pose. En effet, le nombre de particules augmente exponentiellement par rapport aux degrés de liberté du modèle de la personne. Deux caractéristiques influencent grandement les résultats. La première est le nombre de particules générées à chaque partition, qu'on peut varier pour améliorer de manière significative le temps de calcul au détriment de la précision de la pose estimée. La deuxième est représentée par les fonctions de vraisemblance utilisées qui permettent de donner un poids à chaque particule, ce qui est crucial lors de l'utilisation de filtre à particules. Des améliorations du filtre à particules ont été proposées par Deutscher et al dans
[Deutscher et al., 2000], [Deutscher and Reid, 2005], avec la mise en place de "l'Annealed Particle Filtering" ou filtre particulaire à recuit simulé (APF). Ce dernier est utilisé pour estimer la position d'une personne sans marqueur dans un système multi-caméras. Elles combinent un recuit simulé avec un filtre à particules qui s'avère efficace pour l'estimation

de la pose d'un coprs articulé, dans une configuration avec beaucoup d'inconnus. Le filtre à particules traditionnel a cependant un inconvénient, du fait qu'il peut facilement être attiré par des maximums locaux. Dans le filtre à particules recuit, l'ensemble de particules éparses est capable de se déplacer progressivement vers le maximum global sans être distrait par les particules locales. Deutscher [Deutscher and Reid, 2005] introduit également un opérateur de « crossing-over » (similaire à celui que l'on trouve dans les algorithmes génétiques) dans le cadre du filtrage particulaire. Il démontre, que cet opérateur améliore le système de suivi 3D dans les espaces de recherche, dans le cas des objets articulés. On peut citer l'utilisation du filtre particulaire à recuit simulé dans les travaux de Sigal et Balan [Sigal et al., 2010] qui présentent l'ensemble de données HumanEva [1] pour l'évaluation quantitative de méthodes concurrentes d'estimation de pose 3D de personnes. On note cependant, que les méthodes stochastiques posent le problème de la répétabilité du suivi. Pour plusieurs exécutions sur un même jeu de données, les résultats obtenus ne sont pas identiques, de par la nature aléatoire du procédé, contrairement aux méthodes déterministes qui permettent une solution toujours identique pour un problème donné.
Dans le cas des images omnidirectionnelles, peu de travaux se sont penchés sur l'estimation de pose 3D. On peut citer les travaux de [Caron et al., 2009], sur le suivi des objets 3D en utilisant des images stéréoscopiques catadioptriques données par un capteur composé de quatre miroirs et d'une caméra. Le travail présenté repose sur une approche d'asservissement visuel virtuel, adaptée et sur une technique de calcul de pose non-linéaire. Les résultats montrent une robustesse pour les changements d'éclairage, un mauvais alignement et une robustesse encore plus élevée avec quatre miroirs plutôt qu'avec deux. Cependant, cette estimation de pose est présentée pour des objets 3D simples (par exemple une boite).

1. HumanEva est une référence standard pour l'estimation de poses 3D multi-vues

4.1.2 Approche Mono-caméra

L'estimation de la pose de personnes à partir de séquences d'images monoculaires est une tâche encore plus difficile de par le fait que l'information de profondeur n'est plus connue. Dans l'estimation de pose humaine monoculaire, les contraintes cinématiques du modèle humain sont généralement utilisées [Bregler et al., 2004], [Wachter and Nagel, 1997]. Dans le travail de Wachter et Nagel [Wachter and Nagel, 1997], un filtre de Kalman étendu est utilisé pour estimer la posture humaine avec des contraintes cinématiques. Sminchisescu et Triggs [Sminchisescu and Triggs, 2003a], ont étudié l'application de l'échantillonnage stochastique pour estimer la posture humaine monoculaire. Ils utilisent une métrique de coût combinant un flot optique, le contours de la silhouette et les modèles à priori pour l'appariement d'images. L'échantillonnage à covariance est introduit pour guider les particules du filtre et réduire les minimums locaux incorrects. Dans leurs recherches [Sminchisescu and Triggs, 2003b], les auteurs utilisent un raisonnement cinématique simple pour déterminer les potentiels marches vers l'avant ou vers l'arrière (de la personne) qui peuvent provoquer des ambiguïtés visuelles. Des approches probabilistes utilisant des parties du corps humain associées à une cinématique humaine ont également été étudiées pour l'estimation de poses de personnes dans le cas monoculaire. Dans [Lee and Cohen, 2004], les auteurs utilisent une approche de Monte-Carlo par chaînes de Markov (MCMC) afin de modéliser l'estimation de pose des différentes parties du corps dans l'espace 3D avec un modèle 3D explicite. MCMC a été appliquée pour estimer des poses 3D à partir d'images individuelles de joueurs sportifs dans une variété de poses complexes. Néanmoins, cette approche souffre d'un temps de calcul élevé. Plus récemment, dans les travaux de [Yasin et al., 2016], l'estimation de pose 3D avec une seule caméra se fait d'une manière différente. Leur approche repose sur deux sources d'information. La première source est une

base de données de capture de mouvement qui ne contient que des poses 3D. La deuxième source est une base de données d'images avec des poses 2D annotées. Les données de capture de mouvement 3D sont dans un premier temps normalisées et ensuite projetées pour obtenir differentes poses 2D en utilisant plusieurs caméras virtuelles. Cela donne plusieurs paires 3D-2D où les poses 2D servent de caractéristiques pour les poses 3D. La base de données d'images 2D annotées est utilisée afin d'apprendre un modèle de structure picturale (pictorial structure model PSM) [Felzenszwalb and Huttenlocher, 2005] pour l'estimation de pose 2D grâce à l'apprentissage par un classifieur de type forêts d'arbres décisionnels (random forest classifier). Étant donné une image de test, le PSM prédit la pose 2D la plus vraisemblable qui est ensuite utilisée pour récupérer les poses 3D normalisées les plus proches. La pose 3D finale est ensuite modifiée en minimisant l'erreur de projection sous la contrainte que la solution soit proche des poses récupérées. Ces étapes peuvent être répétées pour affiner la pose 3D finale. On peut aussi citer de récents travaux, qui se sont aussi basés sur l'apprentissage de réseaux de neurones comme dans [Nie et al., 2017]. Néanmoins les travaux basés sur les PSM doivent calculer le gradient pour différentes parties du corps afin de pouvoir extraire un HOG pour chaque partie du corps comme dans [Yang and Ramanan, 2013] et [Felzenszwalb et al., 2010b]. Cela nous parait difficilement possible dans le cas des caméras omnidirectionnelles car, même si le descripteur HOG permet la caractérisation de la personne, il est difficilement exploitable pour effectuer un apprentissage sur les différentes parties du corps de celle-ci dans l'image omnidirectionnelle. Pour le cas de l'estimation de pose 3D monoculaires dans les images omnidirectionnelles, à notre connaissance peu de travaux portent sur cette thématique. On peut néanmoins citer [Taiana et al., 2010] qui utilisent les filtres particulaires combinés avec des fonctions de vraisemblance basées sur la couleur et la forme pour faire le suivi 3D d'objets simples, comme une sphère ou un parallélépipède rectangle. D'autres travaux sur l'esti-

mation de pose 3D avec des caméras omnidirectionnelles [Georgakopoulos et al., 2016] et [Delibasis et al., 2016] sont basés sur un apprentissage afin de permettre cette estimation. Néanmoins, il semble que les poses estimées soient limitées et ne permettent pas d'assurer le suivi 3D. Dans le cadre de ce travail de thèse, nous avons proposé un suivi 3D basé sur les filtres particulaires à recuit simulé (APF) avec des contraintes fortes sur le modèle, en mettant en place des fonctions de vraisemblance adaptées que nous présenterons dans la section suivante.

4.2 Méthode pour un suivi 3D stochastique

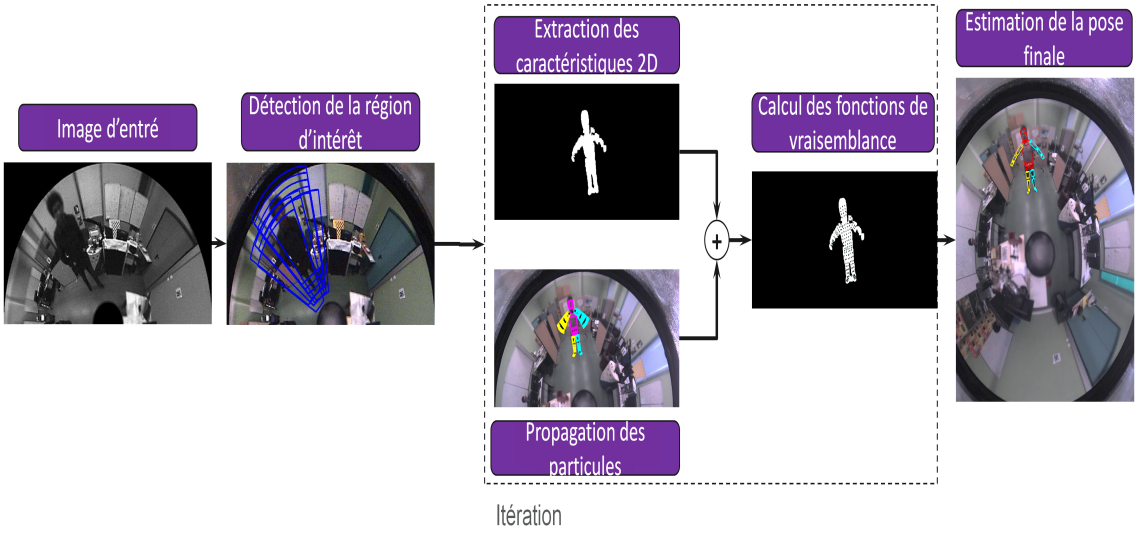


FIGURE 4.1 – Les étapes du suivi 3D

Nous présentons dans cette partie l'approche proposée pour le suivi 3D. Elle comporte plusieurs étapes. La première étape concerne la détection de la région d'intérêt (ROI). Pour cela, nous avons mis en place un algorithme de détection de personnes basé sur les descripteurs HOG dans les images omnidirectionnelles et utilisant le gradient dans des

variétés riemanniennes. Cette étape de détection de la personne qui est traitée dans le chapitre 3, nous permet ainsi de détecter la personne et de réduire la ROI dans l'image omnidirectionnelle dans le cas du suivi 3D. L'initialisation du modèle peut se faire soit automatiquement grâce à cette détection, comme en peut le voir dans la (figure 4.1), soit par l'utilisateur. A partir de la pose initiale, nous pouvons générer plusieurs positions. Chaque position représente une particule respectant les possibilités de mouvement du corps humain qui permettra de déterminer la pose de l'image suivante. Le poids devant être attribué à chaque particule, est calculé suivant une ou plusieurs fonctions de vraisemblance. Notre travail s'est intéressé plus particulièrement à ces fonctions de vraisemblance. Pour tenir compte des distorsions engendrées par notre capteur, nous avons mis en place plusieurs fonctions de vraisemblance adaptées, comme l'utilisation du gradient sphérique pour l'attribution des poids selon la distance géodésique entre le modèle projeté et le contour. Cela permet d'avoir une meilleure estimation du voisinage dans l'espace $\left(S^2\right)$. Nous détaillerons chaque partie dans les sous-sections suivantes.

4.2.1 Modèle 3D

Dans les travaux qui ont porté sur l'estimation de pose 3D, il existe une grande variété de modèles cinématiques, que ce soit par le nombre de degrés de liberté ou par la forme des diférentes parties du corps. Le nombre de degrés de liberté peut varier de 82 ddl, comme dans les travaux de [Yu et al., 2017], à 14 ddl [Azad et al., 2004] [Fontmarty, 2008](14/22 ddl). Dans la majorité des cas 32 ddl sont considérés [Holte et al., 2012]. Le nombre de dégrés de liberté du modèle va représenter le nombre de variables à estimer. Cela necessite de trouver un compromis entre la vitesse de traitement et le nombre de pose (precision) souhaité. Cela dépendra éventuellement du type d'application souhaitée. Afin d'effectuer le suivi 3D d'une personne, nous avons choisi de modéliser le corps suivant un

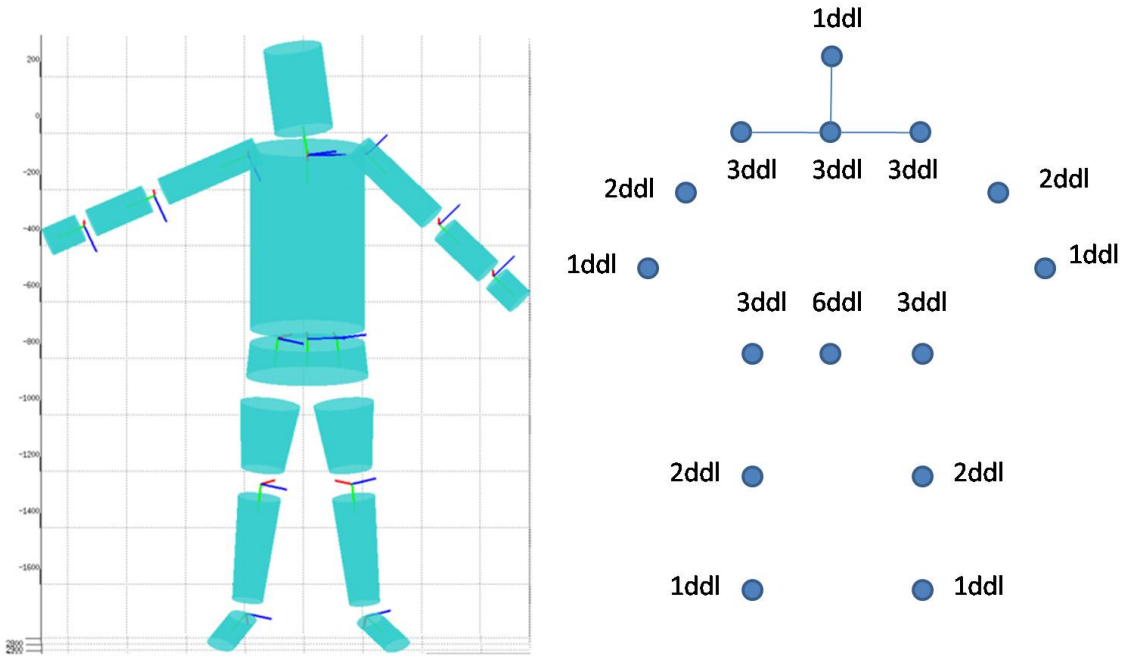


FIGURE 4.2 – Degrés de liberté du modèle 3D

arbre cinématique de 30 degrés de liberté (figure 4.2)qui représente 34 ddl nous negligerons lors des calculs les orientations des mains et des pieds. Les membres sont représentés par des cylindres/cônes tronqués. Cette représentation est assez répandue dans la littérature [Menezes et al., 2006], [Navaratnam et al., 2005] de par la facilité de sa manipulation et de sa projection. Ainsi le modèle est composé de 11 parties : le bassin, le torse, la tête, les bras, les avant-bras , les jambes et les cuisses. Il existe deux types de paramètres qui décrivent la pose et la forme du corps. La forme est donnée par la longueur et la largeur des membres, qui dans notre cas, sont censées être connues. Notre objectif est de récupérer la pose du corps qui est définie par un ensemble réduit de 30 paramètres, y compris la position et l'orientation globale du bassin et les angles articulaires relatifs entre les

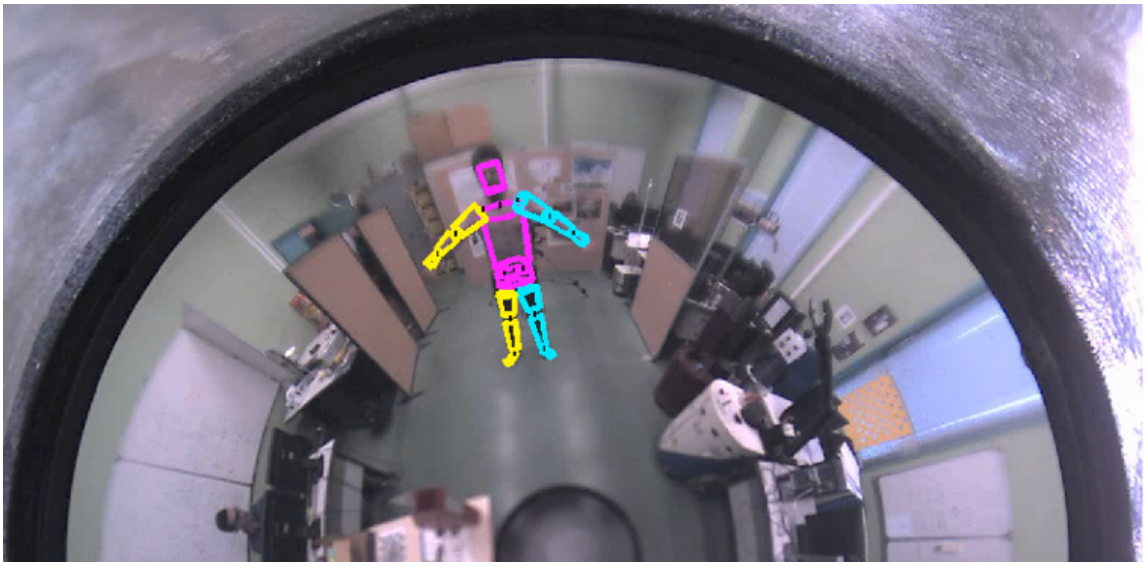


FIGURE 4.3 – Un exemple de la projection du modèle 3D de la personne

membres voisins. Cette configuration permet de représenter le mouvement humain. Ainsi une configuration complète du modèle cinématique est alors donnée par le vecteur d'état $x_t = [x_t^1, x_t^2, ..., x_t^{29}, x_t^{30}]$. Dans notre cas, le vecteur d'état est composé de la position 3D du bassin et de l'orientation des différentes parties du corps par rapport au bassin. Toutes les approches basées sur un modèle doivent intégrer une méthode de projection, généralement basée sur le modèle sténopé. Dans notre cas, la projection est basée sur le modèle unifié, ainsi en projetant les segments de droite situés sur les génératrices de notre cylindre, nous obtiendrons des coniques sur le plan images. Cette projection dans le cas des images omnidirectionnelles peut être plus coûteuse en temps de calcul que dans le cas perspectif. Un exemple de projection des cylindres de notre modèle est illustré dans la (figure 4.3).

4.2.2 Filtre particulaire

Le problème de suivi 3D peut être formalisé dans un cadre stochastique bayésien [Arulampalam et al., 2002] comme une estimation de la distribution de probabilité conditionnelle (dite encore a *posteriori*) $p(x_t|y_{1:t})$. Le vecteur d'état x_t décrit la position du corps à l'instant t ($t \in N$) et $y_{1:t} \equiv (y1, ..., yt)$ représente les observations. Nous nous plaçons dans le cadre d'un système stochastique markovien où la distribution du vecteur d'état à l'instant initial $p(x_0)$ est connue. En considérant un processus de Markov de premier ordre, nous avons $p(x_t \mid x_{1:t-1}) = p(x_t \mid x_{t-1})$ qui représente la dynamique a *priori* du système. En utilisant la règle de Bayes, la distribution de filtrage peut être calculée en deux étapes : (i) étape de prédiction :

$$p(x_n \mid y_{1:n-1}) = \int p(x_t \mid x_{1:t})p(x_{t-1} \mid y_{1:t-1})dx_{t-1} \tag{4.1}$$

(ii) étape de filtrage :

$$p(x_t \mid y_{1:t}) \propto p(y_t \mid x_{1:t})p(x_n \mid y_{1:n-1}) \tag{4.2}$$

où l'intégrale dans (Eq. 4.1) calcule la prédiction en utilisant le modèle de diffusion temporelle $p(x_t \mid x_{t-1})$. La prédiction est pondérée par la probabilité $p(y_t \mid x_t)$ (fonction de vraisemblance) de la nouvelle image d'observation conditionnée par la pose estimée. La solution du problème de filtrage abordée ici peut reposer sur plusieurs méthodes. Le filtre de Kalman [Kalman, 1960] fut le premier utilisé. Toutefois, bien qu'il décrive la solution exacte par des calculs analytiques, il impose des hypothèses contraignantes : la dynamique et le lien état-mesure ($p(y_t \mid x_t)$) doivent pouvoir se modéliser comme des fonctions linéaires auxquelles sont ajoutés des bruits gaussiens. Dans notre cas, nous nous sommes orientés vers une approche par affinement avec l'utilisation du filtre particulaire à recuit simulé (APF). Les méthodes d'approximation non paramétrique représentent les distributions à

posteriori d'un ensemble de N échantillons aléatoires ou particulaire. Les poids normalisés associés à ces particules qui se propagent dans le temps en utilisant le modèle temporel, sont assignés selon une fonction de vraisemblance. l'APF est basé sur les algorithmes de rééchantillonnage d'importance séquentielle (Sequential Importance Resampling SIR), ou sur l'algoritme de CONDENSATION pour « CONDitional DENSity propagATION » de [Arulampalam et al., 2002] [Isard and Blake, 1998]. Ainsi, le filtre à particules recuit (APF) est une variante de SIR. Il est introduit pour le suivi de personnes par Deutscher et Rei [Deutscher and Reid, 2005]. l'APF fait des itérations à chaque instant t afin de mieux localiser le maximum de la fonction de vraisemblance. Il s'appuie sur un recuit simulé pour éviter les optimums locaux. A chaque instant, l'algorithme APF itère dans un ensemble d'étapes (Layer), de la couche m vers la couche 1.Concrètement dans une même image l'APF calcul les Points associés à toutes les particules il va sélectionner la particule le poids le plus important et faire un réechantionnage plus fin on recalculer le poids des nouvelles particules ainsi cette opération sera itéré m fois. Ce qui met à jour la densité de probabilité par rapport aux paramètres d'état. La densité d'état à la couche $m+1$ est représentée par un ensemble de N particules avec leurs poids normalisés associés :

$$S_{t,m+1} \equiv \left\{ x^i_{t,m+1}, \pi^i_{(t,m+1)} \right\}_{i=1}^{N} \tag{4.3}$$

Ou $\pi^i_{(t,m+1)}$ et $x^i_{t,m+1}$ sont respectivement l'état du modèle et son poids associé à un instant t et pour la couche m. Pour l'étape de prédiction de la couche m, un modèle de diffusion gaussien est implémenté. Plus précisément, des hypothèses sont établies avec un rééchantillonnage en utilisant la méthode de Monte-Carlo à partir de la densité de probabilité d'état dans la couche précédente $m+1$ en utilisant :

$$\left\{ x^{(i)}_{t,m} \right\}_{i=1}^{N} \sim \sum_{j=1}^{N} \pi^{(j)}_{t,m+1} \mathcal{N}(x^{(j)}_{t,m+1}, \alpha^{M-m} C) \tag{4.4}$$

Avec C la matrice diagonale de covariance. Cette matrice dépend de la largeur de la recherche à chaque étape. Plus Σ est grand, plus l'espace d'échantillonnage des particules sera large. D'étape en étape, nous pondérons Σ par un paramètre α. Ce paramètre est utilisé pour réduire progressivement la matrice de covariance aux étapes inférieures afin de conduire les particules vers le maximum global de la fonction de vraisemblance. Les poses échantillonnées qui dépassent les limites d'angle articulaire ou qui entraînent une inter-pénétration (position impossible physiquement) des membres, sont rejetées et ne sont pas ré-échantillonnées dans l'étape suivante. Pour vérifier si les poses sont physiquement possibles. Nous avons utilisé la bibliothèque de humaneva [Sigal et al., 2010] qui permet de trouver les différents angles angle limite des mouvements d'une personne. Puis nous effectuons deux vérifications. Dans la première nous vérifions que les avant-bras et le torse ne s'intersectent pas. La deuxième vérification concerne les jambes. Les autres partis du corps ne chevauchent pas, car les angles limites mises en place ne le permettent pas. Les particules restantes reçoivent de nouveaux poids normalisés basés sur une version "recuite" de la fonction de vraisemblance :

$$\pi_{t,m+1}^{(i)} = \frac{p(y_t \mid x_{t,m}^{(i)})^{\beta^m}}{\sum_{j=1}^{N} p(y_t \mid x_{t,m}^{(j)})^{\beta^m}}, i \in 1, ..., N \tag{4.5}$$

où β^m est le deuxième paramètre d'optimisation introduit par Deutscher de telle sorte qu'environ la moitié des particules, soit sélectionnée pour la propagation / diffusion à la couche suivante par un échantillonnage basé sur la méthode de Monte-Carlo (eq. 4.3). L'ensemble de particules résultant $S_{t,m} \equiv \left\{ x_{t,m}^{(i)}, \pi_{t,m}^{(i)} \right\}_{i=1}^{N}$ est ensuite utilisé pour calculer la couche $m-1$ en réappliquant (eq. 4.3), (eq. 4.4) comme on peut le voir dans l'algorithme 1.

On peut noter que SIR est un cas particulier d'APF qui correspond à l'étape de l'AFP ($M = 1$) et pour laquelle l'effet du paramètre ($\beta^m = 1$) est supprimé. l'AFP s'avère

Algorithm 1 Algorithme de l'Annealed Particle Filter.

$\left\{x_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N}$ = APF

if $t = 0$ **then**

Échantillonner $x_0^{(1)}, ..., x_0^{(N)}$ i.i.d. selon $p(x_0)$

et poser $\pi_0^{(i)} = \frac{1}{N}$

end if

if $t \geqslant 1$ **then**

$(x_{t-1}^{(i)}, w_{t-1}^{(i)})_{i-1}^{(N)}$ représente $p(x_{t-1} \mid z_{1:t-1})$

Poser $(x_{t-1}^{(i)}, \pi_{t-1}^{(i)})_{i-1}^{(N)} = (x_{t-1}^{(i)}, \pi_{t-1}^{(i)})^{(N)}$

end if

for $l = L, ..., 1$ **do**

for $i = 1, ..., N$ **do**

Échantillonner indépendamment $x_{t,l}^{(i)} \sim p_l(x_{t,l} \mid x_{t,l+1}^{(i)})$

Associer le poids $\pi_{t,l}^{(i)} \propto \pi_{t,l+1}^{(i)} p_l(z_t \mid x_{t,l}^{(i)})$

end for Normaliser les poids de sorte que $\sum_{i=1}^{N} \pi_{t,l}^{(i)} = 1$

if $l < L$ **then**

Rééchantillonner $\left\{x_{t,m+1}^{i}, \pi_{(t,m+1)}^{i}\right\}_{i=1}^{N}$

end if

end for

affecter $\left\{x_{t,m+1}^{i}, \pi_{(t,m+1)}^{i}\right\}_{i=1}^{N}$

extrêmement efficace. Néanmoins, le choix des paramètres α et β reste difficile. On peut utiliser des heuristiques pour les définir comme dans [Deutscher and Reid, 2005]. Nous avons choisi de les définir selon les essais effectués en prenant en compte les recommandations de [Deutscher and Reid, 2005] qui préconisent un choix pour $\alpha = 0.5$. Dans la section suivante nous présenterons les fonctions de vraisemblance utilisées par notre filtre à particule.

4.2.3 Fonctions de vraisemblance

Les fonctions de vraisemblance permettent l'attribution d'un poids selon l'adéquation entre la projection du modèle à l'instant t et l'état de la silhouette. De nombreuses caractéristiques de l'image pourraient être utilisées, comme les caractéristiques basées sur le flot optique (ou défilement visuel), la couleur et l'apparence. Cependant, les approches les plus courantes sont basées sur les informations de silhouette et de contour. Dans cette section, nous exposerons trois fonctions de vraisemblance mises en place durant cette thèse. Les deux premières fonctions de vraisemblance sont basées sur le contour de la personne en calculant le gradient dans les images omnidirectionnelles et sphériques, ainsi qu'en mesurant la distance géodésique afin de déterminer la distance entre chaque pixel de l'image et le contour. La troisième fonction de vraisemblance est basée sur les silhouettes.

4.2.3.1 Fonctions de vraisemblance basées contours

La fonction de vraisemblance basée sur le gradient est largement utilisée dans les images perspectives. Nous proposons de l'étendre aux cas des images omnidirectionnelles. Nous avons mis en œuvre différentes méthodes pour calculer le gradient dans les images omnidirectionnelles, soit en utilisant une métrique riemannienne, soit en calculant directement le gradient dans l'image sphérique (θ, ϕ).

Dans le cas des images omnidirectionnelles, nous utilisons comme présenté dans le chapitre 2, l'opérateur différentiel sur la variété riemannienne pour calculer le gradient de l'image.

$$g^{ij} = \gamma \begin{pmatrix} -x^2(\xi-1)+\xi+1 & xy(\xi-1) \\ xy(\xi-1) & -y^2(\xi-1)+\xi+1 \end{pmatrix} \tag{4.6}$$

avec :

$$\gamma = \frac{\left(x^2+y^2+(1+\xi)^2\right)^2}{(1+\xi)\left(\xi+\xi^2+\sqrt{1-(x^2+y^2)(\xi^2-1)+2\xi+\xi^2}\right)^2} \tag{4.7}$$

De même que pour le gradient dans les images sphériques, il sera calculé comme montré dans le chapitre 2. Ainsi, le gradient sphérique est défini par :

$$\nabla_{S^2} I_s(\theta,\phi) = \frac{\partial I_s(\theta,\phi)}{\partial\theta} e_\theta + \frac{1}{sin\theta}\frac{\partial I_s(\theta,\phi)}{\partial\phi} e_\phi \tag{4.8}$$

θ et ϕ sont respectivement les angles de longitude et de latitude.
Une fois que le gradient est calculé, la distance entre la projection du modèle dans l'image sphérique et le contour peut être obtenue. En ce qui concerne les images en perspective, la carte des distances est obtenue grâce à la distance de Chanfrein [Barrow et al., 1977]. Elle permet d'estimer la distance euclidienne entre deux pixels dans l'espace discret on associant un coût aux déplacements directs et diagonaux, grâce notamment aux différents masques de chanfrein présentés dans la (figure 4.4).

Cela est possible dans les images perspectives car la distance entre un pixel et son voisinage est toujours identique, quelque soit la position du pixel dans l'image. Or cela n'est pas le cas dans les images omnidirectionnelles comme on peut le constater dans la figure (figure 4.5). Dans le cas des images omnidirectionnelles contrairement aux images

1	1	1
1	0	1
1	1	1

4	3	4
3	0	3
4	3	4

	11		11	
11	7	5	7	11
	5	0	5	
11	7	5	7	11
	11		11	

FIGURE 4.4 – Différents masques pour le calcul de la distance de chanfrein

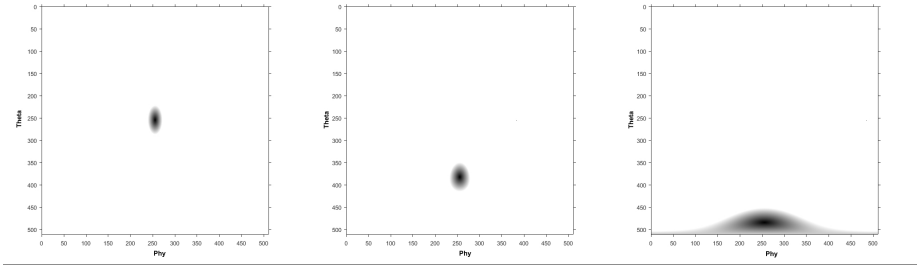


FIGURE 4.5 – Voisinage d'un pixel selon ses coordonnées (θ, ϕ). Dans les trois cas la distance maximale entre le pixel central et le pixel de bord est la même.

perspectives, la distance entre un pixel et son voisinage dépend de la position du pixel dans l'image. Ainsi pour une distance donnée, le voisinage d'un point $\theta = \pi/2$ (figure 4.5(a)) est plus petit que pour une position $\theta = 3 * \pi/4$ (figure 4.5(b)). L'utilisation des distances de chanfrein nous semble inappropriée dans le cas des images omnidirectionnelles. Pour pouvoir calculer la carte de distance nous nous sommes orientés vers l'utilisation de la distance géodésique.

Soit P la projection qui transforme une image omnidirectionnelle R^2 en une image sphérique équivalente. On peut définir la distance géodésique pour deux points x_1 et x_2 dans S^2 avec $x_1 = (\theta_1,\phi_1)$ et $x_2 = (\theta_2,\phi_2)$ comme suit :

$$d_{S^2}(x_1, x_2) = arcos\left(\begin{bmatrix} cos(\phi)sin(\theta) \\ sin(\phi)sin(\theta) \\ cos(\theta) \end{bmatrix} . \begin{bmatrix} cos(\phi_1)sin(\theta_1) \\ sin(\phi_1)sin(\theta_1) \\ cos(\theta_1) \end{bmatrix}\right)$$

Ainsi d_{S^2} représente la distance géodésique.

On peut calculer M_t^e la carte de la distance du contour à l'instant t. La probabilité de vraisemblance est ensuite estimée en projetant dans la carte du contour le modèle des limites apparentes du contour des différentes parties du modèle et en calculant l'erreur quadratique moyenne (MSE) :

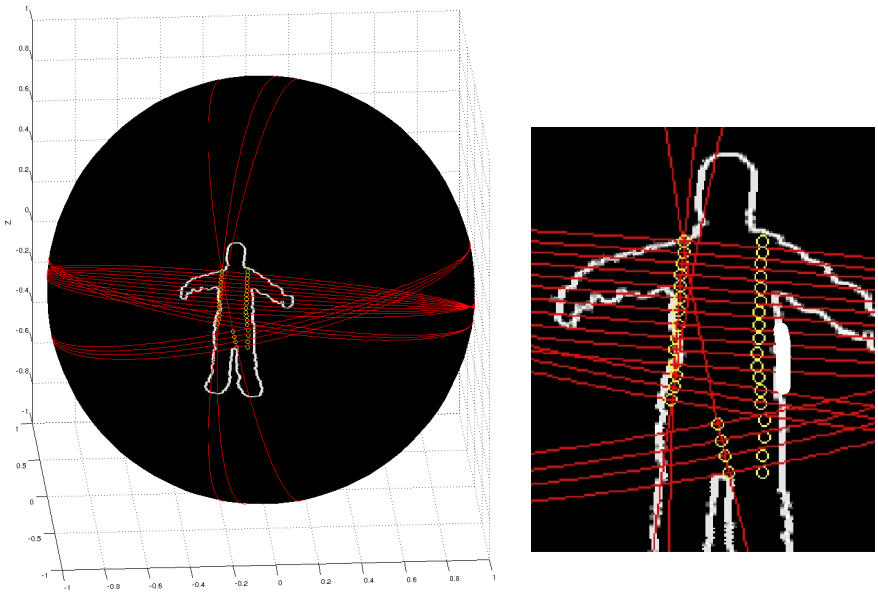
$$P^e(y_t|x_t) \propto \frac{1}{\xi^e_{x_t}(j)} \sum_j (1 - M_t^e(\xi^e_{x_t}(j)))^2$$

où $\xi^e_{x_t}(j)$ est l'ensemble des emplacements de pixels correspondant à tous les points projetés le long de tous les bords des différentes parties du corps, induits par la pose x_t. Pour améliorer la vitesse de calcul, nous avons mis en place une fonction de vraisemblance qui calcule la distance géodésique dans une direction donnée. Pour chaque cylindre du modèle, nous déterminons le grand cercle C1 passant par ses extrémités, puis nous calculons plusieurs cercles qui appartiennent à un plan perpendiculaire à C1 passant par plusieurs points du bord de notre cylindre, comme on peut le voir dans la figure (figure 4.6(b)). La figure (figure 4.6(a)) représente un cercle dans l'image (θ,ϕ). Cela permet de réduire les pixels dont on doit calculer la distance par rapport au contour, car contrairement au cas perspective, la complexité du calcul de la carte de distance est extrêmement importante.

4.2.3.2 Fonctions de vraisemblance basées silhouettes

Fonctions de vraisemblance basée silhouette

Les fonctions de vraisemblance basée sur la silhouette sont comme leur nom l'indique,



(a) Repérsentation sphérique de l'image

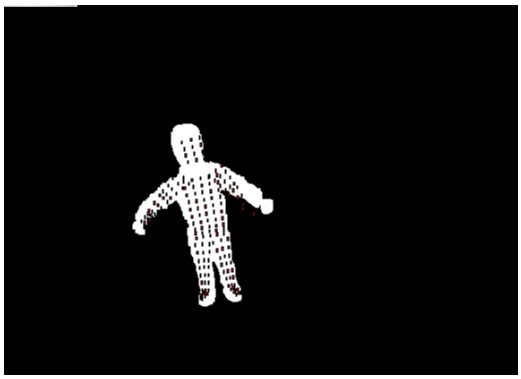
(b) Zoom de l'image (a) 3 parties du corps représentées (a)

FIGURE 4.6 – Représentation de la fonction de vraisemblance basée sur les distances géodésiques dans l'image sphérique

basées cette fois sur la silhouette de la personne et non sur le contour comme dans le cas du gradient. Nous utilisons notre modèle de projection afin de projeter notre modèle 3D échantillonné sur la sphère unitaire. On détermine, l'arrière-plan par soustraction de l'arrière-fond, cela est aisé dans notre cas car la caméra est statique. On utilise pour cela les modèles de mélanges gaussien. Nous définissons une probabilité de silhouette symétrique comme dans [Deutscher and Reid, 2005]. Nous pouvons ainsi générer la carte de silhouette, en soustrayant l'arrière-plan à chaque instant t :

$$P^s(y_t|x_t) \propto \frac{1}{\xi^f_{x_t}(j)} \sum_j (1 - M^s_t(\xi^s_{x_t}(j)))^2$$

(a) Image omnidirectionnelle



(b) Silhouette de la personne dans laquelle le modèle projeté est échantillonné.

FIGURE 4.7 – Fonction de vraisemblance basée silhouette.

Ainsi M_t^s est la carte de silhouette (binaire) à l'instant t, contrairement à M_t^e qui représente la carte de la distance du contour à l'instant t. Dans le cas de la silhouette, il suffit de faire la somme du nombre de pixel appartenant à l'observation de silhouette. Le principal avantage de cette fonction de vraisemblance est la rapidité avec la quelle elle peut être calculée, bien qu'elle impose que le modèle soit toujours projeté à l'intérieur de la silhouette comme dans l'exemple de la figure (figure 4.7).

Fonctions de vraisemblance basée silhouette duale.

La mise en place d'une fonction de vraisemblance basée sur les silhouettes duale permet d'éviter que la silhouette classique qui force le modèle à être à l'intérieur ne finisse par avoir un impact négatif sur le suivi 3D. Nous avons donc mis en place une fonction de vraissemblance comme dans [Sminchisescu and Telea, 2002], qui permet de pénaliser les régions qui ne se chevauchent pas. Nous définissons M_t^p la carte de silhouette binaire de

la projection du modèle et M_t^s la carte de silhouette. Nous avons définie trois régions : R1, R2 et R3 avec R1 l'intersection de M et S, comme on peut le voir dans (la figure 4.8). L'objectif ici est de maximiser la région R1. La taille de chaque région peut être calculée en additionnant tous les pixels de l'image i comme suit :

$$R1_t = \sum_j (M_t^p(i) M_t^s(i))$$
$$R2_t = \sum_j (M_t^s(i)(1 - M_t^p(i)))$$
$$R3_t = \sum_j (M_t^p(i)(1 - M_t^s(i)))$$

Ainsi la fonction de vraisemblance duale est définie comme suit :

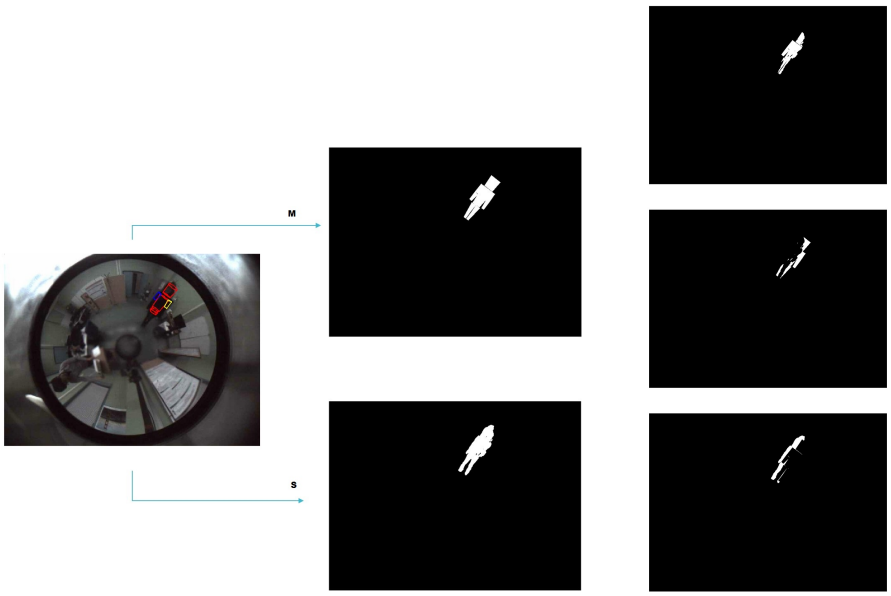
$$P^{sd}(y_t|x_t) \propto \frac{1}{2}(\frac{R2_t}{R1_t + R3_t} + \frac{R3_t}{R1_t + R3_t})$$

4.2.3.3 Combinaison des fonctions de vraisemblances

Plusieurs fonctions de vraisemblance peuvent être mises en place. On suppose que toutes ces fonctions de vraisemblance sont indépendantes entre elles conditionnellement à la position x_t. Cette hypothèse d'indépendance conditionnelle permet de fusionner les mesures potentiellement hétérogènes. Ainsi on peut combiner les mesures de l'image à partir de la formulation à probabilité multiple comme suit :

$$P(y_t|x_t) = \frac{1}{K|L|} \sum_{l \in L} (-logP^l(y_t|x_t))$$

où y_t est l'image à l'instant t et $L \in \{e, s, sd\}$, avec $\{e, s, sd\}$ l'ensemble des fonctions de vraisemblance. Ces différentes fonctions de vraisemblance sont utilisées pour pondérer la



(a) projection du modèle : M et extraction de la silhouette : S (b) (Combinaison S et M, S sans M, M sans S,

FIGURE 4.8 – Fonction de vraisemblance basée silhouette duale

position du modèle à un instant t determinant ainsi le comportement du filtre. Le protocole expérimental mis en place et les résultats préliminaires obtenus sur séquences réelles sont présentés dans la section suivante.

4.3 Résultats

Dans cette section, nous présentons les performances de l'algorithme de suivi 3D implémenté sur des données réelles. La vérité de terrain est ici obtenue à l'aide d'un système Smart-track commercial. Nous détaillons dans un premier temps, le protocole expérimental mis en place, ainsi que la création de notre base de données de test, puis dans une deuxième partie, nous expliquerons les critères d'évaluations retenus.

4.3.1 Protocole expérimental

L'évaluation d'un système de suivi 3D de personnes, nécessite une base de données permettant une évaluation quantitative des algorithmes proposés. L'évaluation de la position 2D des articulations peut être générée avec soin à l'aide d'outils de création de vérité de terrain tels que ViPER [Mariano et al., 2002]. Il est aussi possible de créer cette vérité de terrain 2D de manière manuelle, en annotant les différentes positions des articulations sur chaque image de l'objet à suivre, comme dans les travaux de [Cielniak et al., 2005] ou [Cielniak et al., 2005]. Néanmoins, cela demeure particulièrement long et fastidieux. Dans le cas de l'évaluation de la position 3D, ce n'est pas possible de produire manuellement une vérité 3D à partir de séquences vidéo. De plus, la mise en place de ce type de base de données nécessite un système complexe et coûteux. Dans le contexte des images perspectives, il existe plusieurs bases de données publiques permettant de remédier à cette difficulté. Les bases de données les plus populaires sont CMU [Carnegie-Mellon,] et HumanEva [Sigal et al., 2010]. Chaqu'une d'entre elles, présente différentes séquences avec des difficultés graduelles. Ces bases de données ont permis à la communauté de vision d'utiliser une base commune de comparaison des algorithmes mis en place. Elles ont été utilisées dans plusieurs travaux pour l'évaluation de l'estimation 3D [Zhang and Fan, 2010] [Bo and Sminchisescu, 2010] [Peursum et al., 2010]. Or dans le domaine de la vision non conventionnelle, aucune base de données, similaire n'est proposée. Cela nous a donc poussé à mettre en place une base de données composée de plusieurs séquences que nous présenterons dans les sections suivantes. Une autre difficulté est la comparaison des résultats obtenus à la vérité de terrain, qui nécessite la mise en place de méthodes d'évaluation. Nous présenterons par la suite les différentes méthodes de comparaison mises en place.

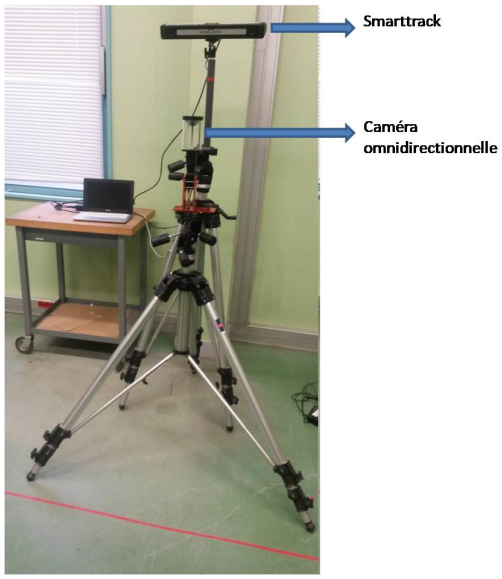


FIGURE 4.9 – Système d'acquisition composé d'un capteur Smarttrack et d'une caméra omnidirectionnelle.

4.3.1.1 Description du dispositif

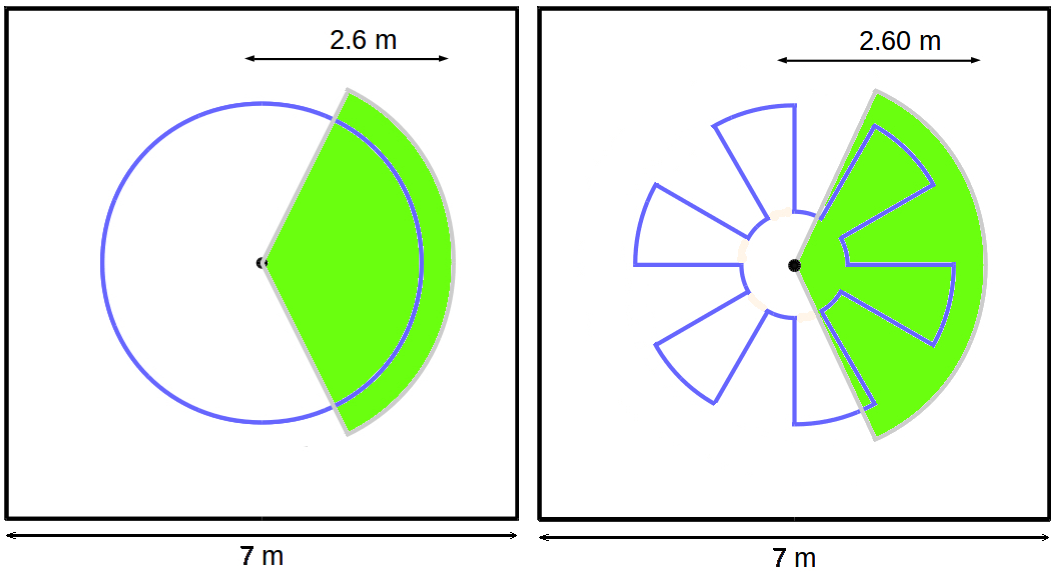
La mise en place d'une base de données de vérité terrain a nécessité un système composé d'un capteur Smarttrack et d'une caméra omnidirectionnelle (un miroir hyperbolique combiné avec une caméra perspective) comme illustré dans la (figure 4.9). Le capteur Smarttrack est composé de deux caméras infrarouges avec un champ de vision pour chaque lentille d'approximativement 100 degrés horizontal et 84 degrés vertical. Avec une fréquence maximale de 60 images par seconde, les caméras infrarouges permettent le suivi de marqueurs composés d'une surface réfléchissantes. Il existe deux types de marqueurs, les marqueurs simples permettant l'estimation de la position 3D et les marqueurs complexes composés de plusieurs marqueurs simples qui permettent l'estimation de la position et l'orientation de l'objet marqué. Les marqueurs complexes seront utilisés pour le cas du bassin et de la

tête. Ils permettront de connaître la position du bassin ou de la tête mais également son orientation. Une tâche difficile mais indispensable consiste, à synchroniser les données entre le capteur omnidirectionnel et le capteur Smarttrack qui permettent de faire correspondre les positions 3D obtenues à l'image acquise. Comme le Smarttrack possède une entrée de synchronisation, et non pas la caméra omnidirectionnelle, nous avons opté pour une synchronisation logiciel, grâce à un serveur WIA [2] qui permet de capturer le flux entrant de la caméra omnidirectionnelle et donc dater les images. Un autre processus gère l'entrée du flux (daté) parvenant du capteur Smarttrack. Ce capteur a une fréquence de 100Hz, or la fréquence de la caméra varie de 6Hz à 20Hz selon la taille des images. Nous avons fait correspondre chaque image aux données du smarttrack correspondantes, avec une précision de synchronisation de moins de 25 millisecondes. Cela nous a permis, de générer notre base de données de test 3D composée d'images omnidirectionnelles annotées. Il faut noter que durant la phase d'acquisition, des artefacts peuvent apparaître. Cela est dû à la présence de surface réfléchissante. Pour remédier à cela, les données ont subi un post traitement permettant de ne conserver que les points 3D proches de la position d'un autre point 3D dans l'image précédente.

4.3.1.2 Génération de la base de données

Grâce au dispositif mis en place, on a pu créer une base de données 3D composée de 4 séquences. La première séquence représente une personne qui se déplace autour du capteur omnidirectionnel figure (figure 4.10). La seconde représente une personne qui se déplace autour de la caméra omnidirectionnelle avec un mouvement oscillant des bras. Dans la troisième séquence, un déplacement autour du capteur avec un déplacement avant/arrière en direction du capteur à été effectué. La quatrième séquence est réalisée avec une difficulté

2. Un logiciel d'acquisition d'images



(a) Séquence 1 et 2 (b) Séquence 3

FIGURE 4.10 – Trajectoire de déplacement sur les trois séquences. En vert le champ de vision du capteur Smartrack, le capteur omnidirectionnel a un champ de vision qui recouvre toute la zone de déplacement. (a) séquence 1 et 2 déplacement (bleu) circulaire autour du capteur, avec des mouvements des bras pour le cas de la séquence 2. (b) Déplacement circulaire avec mouvement avant/arrière pour la séquence 3.

accrue pour le suivi 3D. Elle contient une rotation de la personne sur elle-même et une montée de marche permettant d'évaluer la robustesse de l'algorithme face à une auto-occultation. L'une des contraintes générées par l'utilisation du capteur Smarttrack est la relative faible distance de détection qui selon le fabricant est de l'ordre prés de 2 m. Nous avons pu constater que cela pourrait aller jusqu'à 2.6 m avec un taux d'erreur acceptable qui peut être corrigé lors du post traitement et qui est de l'ordre de 5 mm. Ceci n'est possible que dans le cas des marqueurs complexes. Cette faible distance de détection du Smarttrack nous a poussé à subdiviser chacune de ces 4 séquences vidéo en 2, comme présenté dans le

	Séquence 1		Séquence 2		Séquence 3		Séquence 4	
	2D	3D	2D	3D	2D	3D	2D	3D
Nombre d'images	380	220	402	280	473	295	140	111
Durée de la séquence	31	18	33	23	39	24	12	9
Type de mouvement	Circulaire		Circulaire + Bras		Avance/Recule		Occlusion/marche	

TABLE 4.1 – Les différentes séquences monoculaires utilisées pour les tests

tableau 4.1. Une partie où l'information de profondeur est connue et l'autre où seule une vérité terrain 2D est possible. Comme on peut le constater dans la figure (figure 4.10), le champ de vision du Smarttrack est relativement restreint comparé à celui d'une caméra omnidirectionnelle. Ainsi dans chaque séquence, seule une partie sera évaluée selon une vérité de terrain 3D. L'autre partie est annotée manuellement pour chaque articulation dans toutes les images où la position 3D n'est plus connue. La vérité terrain 2D, a été mise en place en utilisant plusieurs logiciels d'annotation tels que ViPER ou la Toolbox Computer Vision système de Matlab. Cependant, la tache demeure fastidieuse. Pour le cas des séquences 3D, seule le calibrage entre le Smarttrack et la caméra omnidirectionnelle, est nécessaire.

4.3.1.3 Les critères de performances

Nous souhaitons définir les différentes mesures permettant de comparer les différentes stratégies de filtrage utilisées. Diverses méthodes d'évaluation 2D et 3D ont été proposées dans la littérature, pour évaluer le suivi du mouvement humain et l'estimation de la pose. Par exemple, un certain nombre de travaux suggèrent d'utiliser la différence d'angle des joints comme mesure d'erreur [Ning et al., 2008] [Navaratnam et al., 2007], ou encore la

superposition des pixels (Pixel overlap) entre le résultat obtenu et la vérité terrain, utilisée dans [Srinivasan and Shi, 2007]. Pour nos expérimentations, nous avons décidé de mettre en place deux méthodes de comparaison. Pour l'ensemble de données 3D, nous utilisons comme mesure la racine de l'erreur quadratique moyenne largement utilisée dans la littérature, basée sur l'ensemble des marqueurs virtuels qui correspondent aux emplacements des articulations et des extrémités des membres. Cette mesure d'erreur a d'abord été introduite pour l'estimation et le suivi de la pose 3D dans [Sigal et al., 2004] et étendue dans [Balan et al., 2005]. L'erreur 3D est calculée comme suit :

$$D(x, \hat{x}) = \frac{1}{M} \sum_{i=1}^{M} \|m_i(x) - m_i(\hat{x})\|$$

où $m_i(x) \in R^3$ est la position du marqueur 3D de la pose x. Ainsi, l'erreur 3D représente la distance en (mm) entre notre estimation et la vérité de terrain. La deuxième méthode est basée sur l'erreur 2D entre la projection du modèle et la vérité de terrain directement dans les images omnidirectionnelles. Nous l'avons utilisée sur les séquences vidéo de vérité terrain 2D. Pour obtenir l'erreur 2D comme dans [Lan and Huttenlocher, 2005], nous calculons la distance quadratique moyenne entre les 11 extrémités de notre modèle projeté dans l'image et les points de l'image annotée. La distance obtenue est en pixels :

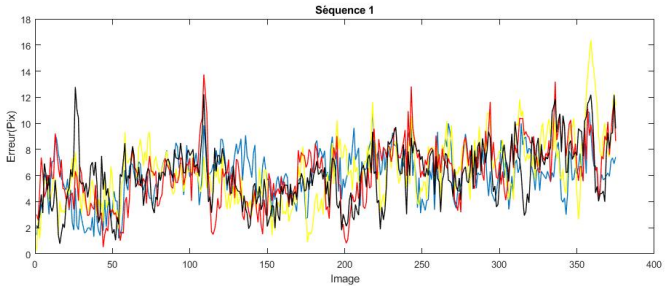
$$D(x, \hat{x}) = \frac{1}{M} \sum_{i=1}^{M} \|d_i(x) - m_i(\hat{x})\|$$

avec $d_i(x) \in R^2$ la projection sur l'image du marqueur 3D de la pose x.

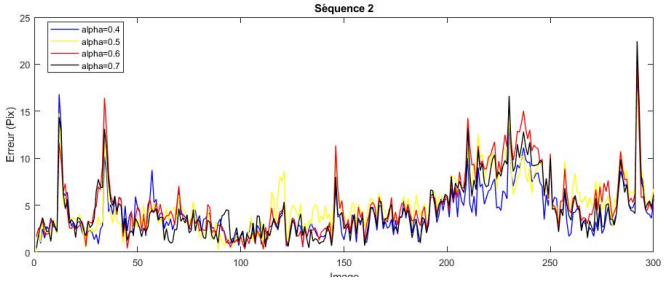
4.3.2 Expérimentations

Nous avons effectué une série d'expériences avec deux méthodes de filtrage APF et SIR, ainsi qu'avec plusieurs fonctions de vraisemblance et différents modèles dynamiques du

mouvement. Les détails de chaque variante sont décrits avec l'expérience correspondante. Les expériences ont été menées sur les trois séquences de l'ensemble de données présentées. Dans chaque cas, la première image pour initialiser le suivi est supposée connue. Nous allons dans un premier temps nous concentrer sur les différents paramètres de notre filtre à particule à recuit simulé permettant un suivi optimal.



(a) Séquence 1

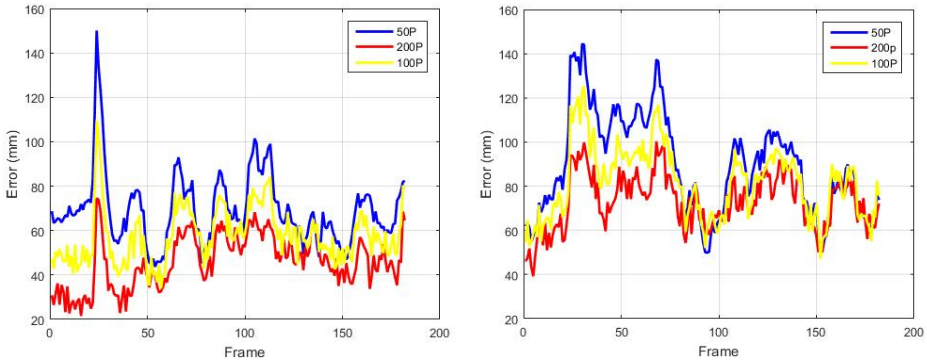


(b) Séquence 2

FIGURE 4.11 – Erreur pixellique montrant l'influence du paramètre α (a) séquence 1 (b)séquence 2

4.3.2.1 Évaluation des paramètres du filtrage particulaire

La stratégie d'optimisation utilisée est de nature stochastique, ce qui peut engendre des résultats différents lors de l'exécution de la même expérience avec les mêmes paramètres de configurations. Ce qui est normal car les particules générées peuvent être différente d'une expérience à une autre. Néanmoins ceci ne permet pas une évaluation sur une seule expérimentation. Afin d'obtenir des mesures cohérentes et une certaine répétabilité des performances, nous exécutons chaque plusieurs fois l'expérience pour chacune des séquences. Nous calculons la moyenne de notre mesure d'erreur (3D ou 2D selon l'expérience) pour chaque instant sur toutes les positions estimées. Dans un premier temps, nous évaluons l'effet du paramètre de rééchantillonnage α du filtre particulaire, utilisé dans l'APF, qui permet de restreindre la propagation des particules de la couche M à la couche M-1. Dans [Deutscher and Reid, 2005], les auteurs préconisent le choix de $\alpha = 0{,}5$. Nous constatons comme on peut le voir sur la figure (figure 4.11), que le paramètre α optimal peut faire varier les résultats obtenus particulièrement lorsque le nombre de particules est peu important. Ainsi, dans les séquences 1 et 2 de la figure 4.11, nous avons fait varier la valeur α de 0,4 à 0,7. On constate que la valeur $\alpha = 0{,}4$ permet d'obtenir les meilleures performances. Cela est dû au fait de la non limitation de l'espace de propagation d'une couche à la suivante de manière trop rapide, particulièrement lorsque les mouvements sont conséquents. C'est le cas des bras dans la séquence 2 où le système ne permet plus de suivre les articulations qui ont subi un grand mouvement. L'erreur pixélique dans le cas de $\alpha = 0{,}4$ est de 6,04 pixels dans la séquence 1 et de 4,15 pixels dans la séquence 2. La valeur $\alpha = 0,6$ donne les résultats les plus médiocres avec une erreur de 6,32 pixels pour la séquence 1 et de 5.06 pixels pour la séquence 2. Ainsi le choix du paramètre α peut dans certains cas, améliorer les performances de détections de 22%, comme le cas de la séquence 2.



(a) Séquence 1 (b) Séquence 2

FIGURE 4.12 – Influence du nombre de particules N (a) séquence 1 (b) séquence 2

Dans la figure (figure 4.12) nous faisons varier le nombre de particules utilisées par le filtre APF. Ce dernier utilise, la même fonction de vraisemblance basée contour et silhouette pour ce test. Nous évaluons l'erreur 3D par rapport au nombre de particules utilisées. Plus le nombre de particules augmente, plus la précision du suivi sera accrue. Nous pouvons constater qu'à partir de 200 particules, l'algorithme donne des résultats corrects et stables. Cependant, lorsque le nombre de particules est inférieures à 100, l'erreur devient plus importante. Il est à noter que le temps de calcul dépend du nombre de particules utilisées. Plus le nombre de particules augmente, plus le temps de calcul augmente. Un compromis entre la vitesse et la robustesse est donc nécessaire. De plus, il a été constaté qu'au-delà de 500 particules, l'augmentation du nombre de particules ne permet plus d'améliorer la détection et le suivi. Au vu du temps d'exécution, important pour une APF de plusieurs couches, il nous a semblé important de pouvoir quantifier l'impact que peut avoir le nombre de couches sur le suivi, même s'il semble évident que plus le nombre de couches sera important, plus le suivi sera performant. Nous évaluons donc le nombre de couches afin de déterminer sa valeur pour laquelle l'amélioration devient négligeable. Comme on peut

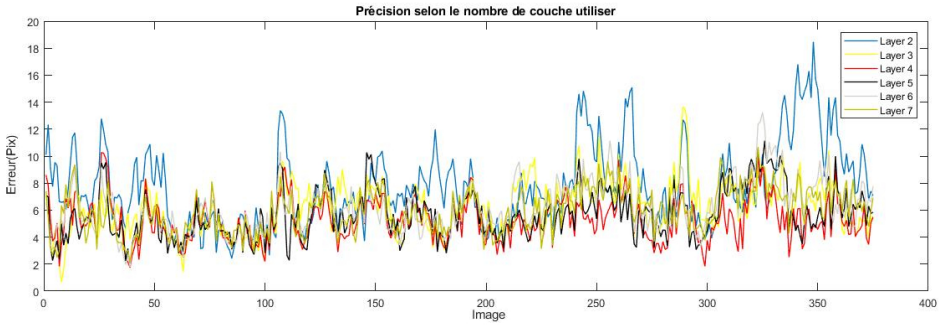


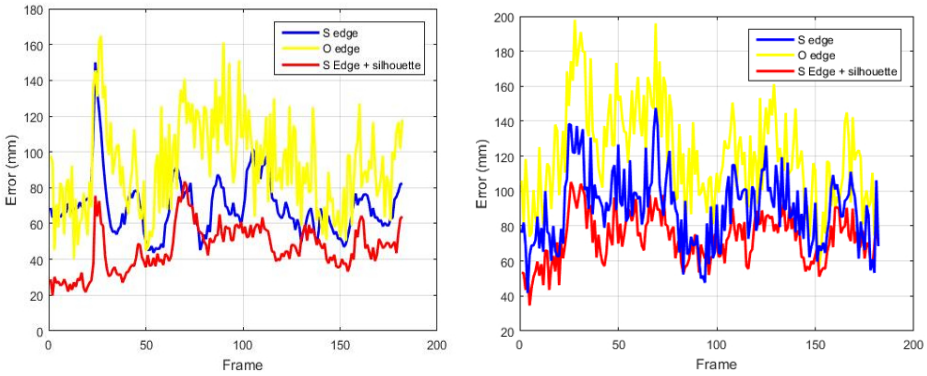
FIGURE 4.13 – Évaluation du nombre de couche séquence 3.

le voir dans la figure (figure 4.13), 4 couches permettent un suivi optimal de la personne. L'augmentation du nombre de couches n'améliore plus les performances de détection de manière significative.

4.3.2.2 Évaluation des fonctions de vraisemblance

En utilisant uniquement la fonction de vraisemblance basée silhouette, notre système de suivi ne parvient pas à effectuer le suivi de la séquence 1 qui est la plus simple, malgré son faible temps de calcul. L'utilisation unique de la silhouette simple ne semble pas un choix pertinent. En comparant les deux fonctions de vraisemblance basées sur le gradient, sur l'espace omnidirectionnel (le Gradient Omnidirectionnel (GO)) et sphérique (Gradient sphérique avec la distance Géodésique (GS)), nous pouvons remarquer que la méthode de gradient sphérique permet d'améliorer les résultats de 11% par rapport à celle basée sur le gradient omnidirectionnel. Cela est dû, au fait que l'image sphérique permet une meilleure représentation des images omnidirectionnelles. Dans l'espace sphérique, le corps a la même dimension pour chaque partie du corps quelle soit sa position dans l'image. De plus, l'utilisation des distances géodésiques semble donner de meilleurs résultats. Cela grâce au fait que la distance calculée entre le contour de la personne extrait et le contour

du modèle projeté est mieux adaptée et donc plus précise. Ainsi cela permet de faire, une discrimination entre deux particules qui dans le cas de l'utilisation de la distance de chanfrein aurait semblé équivalente. Néanmoins, cette précision se fait au détriment d'un temps de calcul plus important.



(a) Séquence 1 (b) Séquence 2

FIGURE 4.14 – Fonction de vraisemblance selon l'erreur 3D

Nous avons également testé les différentes combinaisons de fonctions de vraisemblance GS, GO et la Silhouette Duale (SD), dans le but de déterminer la fonction de vraisemblance la mieux adaptée, permettant d'obtenir un suivi robuste avec des caméras omnidirectionnelles. Nous avons aussi opté pour le choix de la silhouette duale au lieu de la silhouette simple. En effet, prendre en compte un ratio du modèle dans la silhouette et de la silhouette dans le modèle permet d'améliorer le suivi et surtout d'éviter que le modèle ne soit enfermé dans la silhouette. Nous pouvons remarquer dans la figure (figure 4.14) que la combinaison de SD et GS permet d'obtenir de meilleurs résultats. Cette combinaison est plus performante que l'utilisation du gradient sphérique uniquement. Cela est dû à l'apport d'information qu'apporte la silhouette et la projection du modèle dans la silhouette, ce qui permet de garder le modèle projeté à l'intérieur de la silhouette. Dans la figure (figure 4.15),

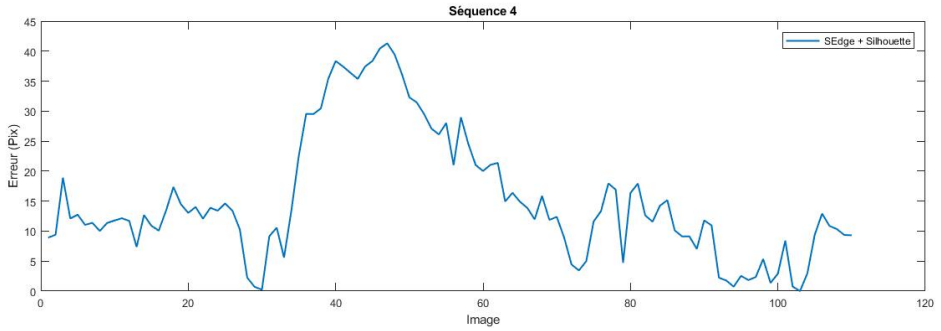


FIGURE 4.15 – Évaluation de la séquence 4.

un test de meilleur paramétrage sur la séquence 4 a permis d'obtenir en moyenne une erreur de 15.2 pixels dans chaque image. La personne dans cette séquence est moins bien suivie que dans les 3 premières séquences. En ce qui concerne les séquences précédentes, l'erreur moyenne est comprise entre 4.58 et 6.7 pixels comme on peut le voir dans le tableau 4.2. Cela s'explique par la difficulté des postures dans la séquence 4 comme l'auto-occultation. Ainsi, lorsque la personne tourne sur elle-même, deux cas de figure sont à distinguer. Soit la personne fait une rotation avec les bras écartés, dans ce cas le suivi est possible avec une bonne robustesse. Soit les bras sont collés au corps, dans ce cas de figure ni le contour ni la silhouette, ne peuvent apporter suffisamment d'information pour détecter cette rotation de la personne. Par conséquent, les fonctions de vraisemblances seules ne permettent pas de détecter cette rotation. Il est possible de recadrer le suivi 3D en mettant en place des fonctions qui peuvent détecter cette rotation par la suite en utilisant la dynamique de mouvement de la personne dans la séquence d'images.

4.3.2.3 Évaluation de la dynamique de mouvement

Rappelons que l'implémentation APF nécessite un modèle de déplacement de la personne afin d'effectuer la propagation des particules. Deux modèles dynamiques $p(x)$, décrivant

	Séquence 1	Séquence 2	Séquence 3	Séquence 4
Méthode Silhouette	56 ± 4.1 pix	X	X	X
Méthode Silhouette Dual (SD)	6.86 ± 0.7 pix	7.15 ± 0.65 pix	7.95 ± 0.76 pix	20.15 ± 1.51 pix
Méthode Gradient sphérique (GO)	6.37 ± 0.6 pix	8.15 ± 0.72 pix	7.01 ± 0.73 pix	22 ± 1.86 pix
Méthode Gradient omnidirectionnel (GS)	4.4 ± 0.45 pix	5.7 ± 0.53 pix	7.2 ± 0.62 pix	18.4 ± 1.63 pix
Méthodes (GS+SD)	4.58 ± 0.42 pix	5.30 ± 0.58 pix	6.72 ± 0.61 pix	15.2 ± 1.26 pix

TABLE 4.2 – Erreur pixelique selon les fonction de vraisemblances utilisées.

la connaissance a priori sur l'évolution temporelle du modèle de l'homme, caractérisée par son état x_k, sont exploités. La première méthode est possible grâce à l'étude d'une première séquence d'apprentissage. Cela permettra de générer une matrice de covariance d'échantillonnage ainsi que la limite des différents angles entre les articulations. Pour le choix de la séquence d'apprentissage, nous avons opté pour une séquence décrivant le plus de mouvements possibles. Cela permet d'avoir une matrice de covariance non seulement d'un mouvement particulier marche, course comme dans [Poppe, 2010], mais également d'un mouvement générique. Les limitations d'angles, seront celles possibles par le corps humain. Le deuxième modèle dynamique ne repose sur aucune connaissance a priori des mouvements effectués. La seule hypothèse raisonnable considère l'état x_k à l'instant k proche de l'état x_{k-1} à l'instant précédent. Ceci est possible grâce à la vitesse d'acquisition du capteur. Ce type de modèles bruités est appelé «marche aléatoire». Il repose sur l'utilisation d'une dynamique gaussienne centrée sur l'état à l'instant précédent :

$$p(x_k|x_{k-1}) = \mathcal{N}(x_k; x_{k-1}, \nabla_k). \tag{4.9}$$

où ∇_k est la matrice diagonale de covariance. Échantillonner $p(x_k \mid p_{k-1})^\alpha$ revient alors à échantillonner $\mathcal{N}(x_k; x_{k-1}, \frac{1}{\alpha}\nabla_k)$. Nous avons défini les écarts types de x_k comme suit : 0.1 m pour les translations et 1.5 degrés pour les rotations, sauf dans le cas de la rotation de l'avant-bras par rapport au torse où l'écart type est plus important (3 degrés). La figure (figure 4.16) montre un exemple de propagation des particules avec les paramètres cités. Ainsi à partir de la position t-1 on généré les nouvelles positions. On remarque que le mouvement des bras et des jambes et plus important que celui du torse. La figure (figure 4.17), illustre les deux modèles dynamiques basés sur la personne. Il nous est apparu, que l'utilisation de la marche aléatoire comme modèle de propagation permet d'avoir de meilleurs résultats. Cela semble dû au fait que malgré l'étape d'apprentissage, le modèle générique ne permet pas dans certains cas, comme dans l'exemple de la séquence 2, de suivre correctement le mouvement des bras lorsque l'angle entre les bras et le torse est très élevé. Le modèle avec apprentissage ne semble être performant que dans le cas de mouvements appris et reproduits dans la séquence de test. Ceci limite cependant son champ d'application. Nous avons ainsi opté pour ≪marche aléatoire≫ comme dynamique de mouvement qui permet un meilleur suivi de la personne si on n'a pas une connaissance *a priori* du déplacement.

4.3.2.4 Évaluation de l'erreur pour les différentes parties du corps

La figure (figure 4.19), montre l'erreur 2D (en pixel) pour les extrémités du corps. Ces extrémités sont représentées par la tête, les mains et les pieds. Nous constatons que la tête est la partie la mieux détectée du corps, alors que les pieds sont moins bien détectés, car leur position dans les images omnidirectionnelles (près du centre) réduit leur taille. En effet la résolution est plus importante lorsqu'on est proche du centre, ce qui rend leur suivi plus difficile. Dans la littérature, les mains sont moins bien suivies que les jambes à cause de l'occultation due au torse. Dans notre cas, les mains sont mieux suivies que les pieds dans

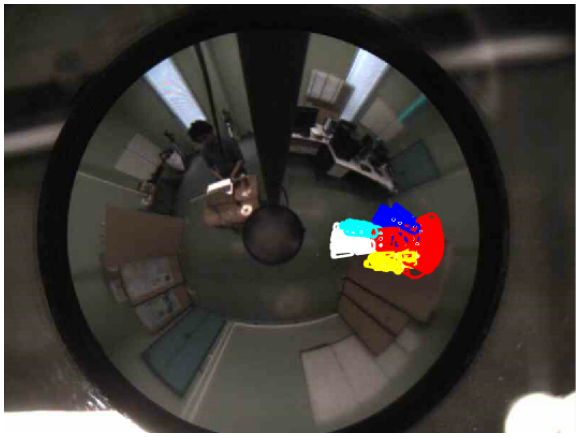


FIGURE 4.16 – Propagation des particules suivant une marche aléatoire.

nos séquences parce que la personne reste toujours face à la caméra. La figure (figure 4.18) illustre un exemple de suivi de la tête par rapport à la vérité terrain.

4.3.2.5 Évaluation du temps de calcul

Le temps de calcul est directement proportionnel au nombre de particules, à la fonction de vraisemblance et au nombre de couches utilisées. Il dépend aussi du choix de la fonction de vraisemblance. Dans le tableau 4.3, les temps de calcul ont été pris pour les cas les plus lents avec l'utilisation d'une combinaison de deux fonctions de vraisemblance (gradient avec distance géodésique et silhouette dual) avec 100 particules pour une seule couche (le temps de propagation et de calcul de la fonction de vraisemblance, sera multiplié par le nombre de couches). Le temps d'exécutions est de 7,35 secondes. La machine utilisée pour les tests est composée d'un processeur Xeon de 3 GHz, l'algorithme est écrit avec Matlab. Il peut facilement être amélioré, néanmoins on peut déjà avoir une idée précise des différentes parties de l'algorithme qui vont demander le plus de ressources et qui devront être optimisées en priorité. Ainsi le temps de pré-traitement des images, du calcul du gradient et de la

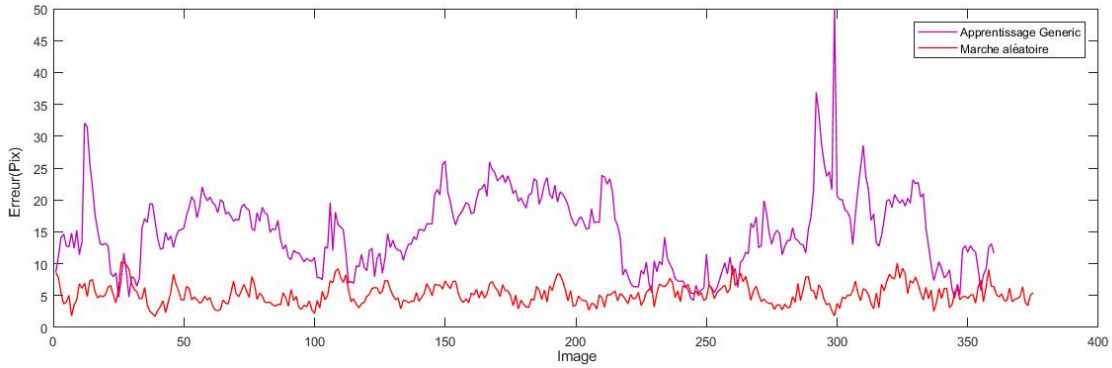


FIGURE 4.17 – Comparaison entre les deux types de modèle dynamique

distance géodésique, représente 57% du temps de calcul global. La soustraction de l'arrière fond, le calcul du gradient sphérique et la distance géodésique ne sont calculés qu'une seule fois par image. Ce temps élevé est justifié par les multiples projections omnidirectionnelles vers l'espace sphérique. De plus, le temps pour calculer la distance géodésique est extrêmement important. Dans notre cas, nous limitons les calculs à un espace restreint de l'image grâce à notre fenêtre de détection HOG. L'évaluation des fonctions de vraisemblance représente 37% du temps de calcul. Pour 100 particules, ce temps de calcul est proportionnel au nombre de couches utilisées. La diffusion des particules et la soustraction de l'arrière fond sont relativement insignifiantes puisque elles représentent 1% du temps de calcul total. L'algorithme APF pour 5 couches, nécessite un temps de traitement de plus de 19,6s qui est beaucoup plus important que celui de SIR (une seule couche) pour un nombre donné de particules. Ceci est facilement compréhensible de par le fait que l'APF effectue les mêmes calculs que le filtre SIR, multiplié par le nombre de couches utilisées. Le temps de calcul peut facilement être amélioré notamment en optimisant les fonctions de calcul du gradient sphérique et de la distance géodésique et en développant l'algorithme en langage bas niveau adapté.

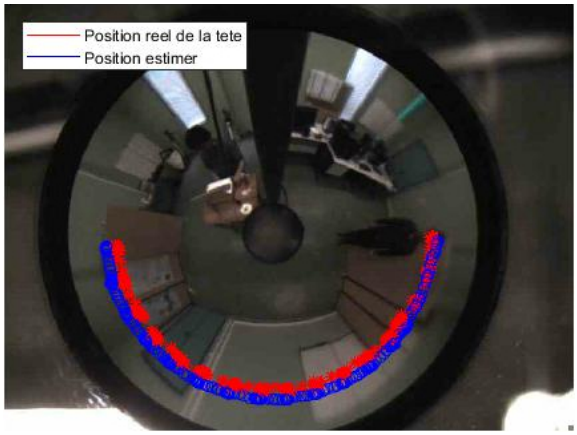


FIGURE 4.18 – Estimation de la position de la tête.

Taille image	800*600	1028*738
Soustraction de l'arrière fond	0.069s (1%)	0,071(1%)
Calculs gradient + distance géodésique	4,2s (57%)	5,01(58%)
Propagation	0,38s (5%)	0,46(5%)
Calculs fonction de vraisemblance (silhouette dual + gradient)	2,7s (37%)	3,15(36%)
Total	7,35s	8,7s

TABLE 4.3 – Temps de calcul des différentes parties de l'algorithme de suivi 3D Pour 100 particules avec m=1

4.4 Conclusion

Nous avons présenté dans ce chapitre plusieurs fonctions de vraisemblance permettant, d'effectuer un suivi 3D de personnes. Le suivi a été réalisé sur des scènes de diverses complexités. Les fonctions de vraisemblance utilisées se basent sur des caractéristiques de silhouette et de gradient. Pour cela, il a été nécessaire d'adapter le calcul du gradient ainsi

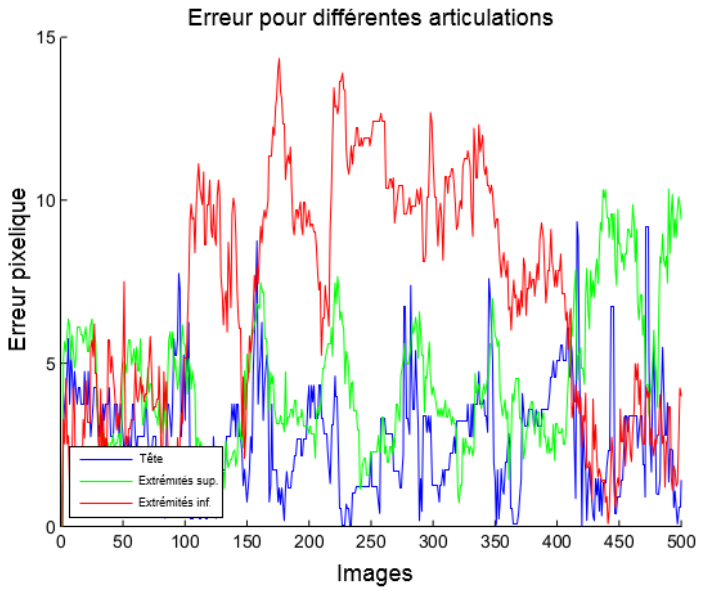


FIGURE 4.19 – Erreur pixelique pour les articulations à l'extrémité du corps.

que la métrique dans l'espace S^2. Ainsi, nous avons utilisé les distances géodésiques définies sur la sphère unitaire afin de générer la carte des distances pour le cas des fonctions de vraisemblance basées sur le gradient. Pour le cas des fonctions de vraisemblance basée sur la silhouette, les projections du modèle se font dans l'espace sphérique de même que l'extraction de la silhouette. Nous avons opté pour une soustraction de l'arrière-plan basée sur un modèle de mélanges de gaussiens (Mixture of Gaussian) plutôt que sur la coupe de graphe (graph cut), car même si la soustraction d'arrière-plan basée sur la coupe de graphe donne de meilleurs résultats, le temps de calcul est beaucoup plus important.

Les résultats montrent que l'utilisation de caméras omnidirectionnelles pour le suivi 3D de personnes peut se faire avec une bonne précision (néanmoins en deçà des précisions obtenues avec des caméras conventionnelles), dès que l'adaptation des algorithmes tient compte de la géométrie du capteur utilisé. Ainsi, l'utilisation du filtre particulaire combiné

à des fonctions de vraisemblance adaptées permettent le suivi 3D monoculaire. Nous avons pu constater que les fonctions de vraisemblance basées sur le gradient sphérique et une distance géodésique combinées avec les fonctions de vraisemblance basées silhouette duale, permettent le suivi 3D avec une meilleure précision. Néanmoins, cela nécessite la mise en place de quelques contraintes qui nous permettent d'estimer l'information de profondeur. Principalement, on suppose que les paramètres de calibration de notre capteur ainsi que la taille des cylindres de notre modèle de la personne sont connus. La partie expérimentale a nécessité la mise en place d'une base de données de vérité terrain 2D et 3D, pour la comparaison des différentes stratégies de suivi 3D. Cela a été rendu possible avec la mise en place d'un système composé d'un Smarttrack et d'une caméra omnidirectionnelle, ainsi que la synchronisation entre les images omnidirectionnelles et les données 3D. Cela, nous a permis la mise en place d'une vérité terrain 3D et 2D précise.

Conclusion générale

Le suivi de personnes est un domaine de recherche extrêmement active de par le nombre d'applications qu'il peut contenir. De nombreuses recherches ont été menées par la communauté sur la détection et le suivi de personnes vues par une ou plusieurs caméras. Ce travail de thèse a pour objectif la valorisation du capteur omnidirectionnel dans cette tâche de détection et de suivi. Ce type de capteur avec son large champ de vision peut, d'une part, limiter la sortie de champ d'une personne et, d'autre part, présenter des résultats comparables à ceux de la litérature lorsque une caméra perspective est considérée. Nous avons donc proposé de mettre en place des algorithmes de détection et de suivi de personnes en utilisant une caméra omnidirectionnelle. Deux approches ont été proposées : 2D et 3D. Nous nous somme intéressés dans un premier temps à la détection et le suivi 2D de personnes dans des images omnidirectionnelles. Nous avons ainsi mis en place une base de données adaptée à la géométrie des caméras omnidirectionnelles. Nous avons ensuite proposé une seconde approche 3D de détection et de suivi de personnes. Nous allons dans un premier temps donner un aperçu des contributions de ce travail. Quelques perspectives de travaux futurs sont ensuite proposées.

Les approches que nous avons proposées dans ce travail de thèse prennent en compte la géométrie de la caméra. En effet, la stéréo-projection sphérique et perspective permet de

modéliser une grande majorité des caméras omnidirectionnelles. Notre première contribution était donc de mettre en place un descripteur adapté aux images sphériques. Nous avons donc adapté l'approche conventionnelle basée sur l'Histogramme Orienté du Gradient pour la détection de personnes dans des images. Notre approche exploite les variétés riemanniennes afin d'adapter le calcul du gradient dans le cas des images omnidirectionnelles. Nous avons ainsi mis en place un descripteur adapté aux images sphériques en utilisant le gradient sphérique. Ces descripteurs sont ensuite combinés avec un classifieur SVM linéaire pour la prise de décision lors du suivi d'une personne dans une image omnidirectionnelle.

Nous avons mis en place pour la phase d'apprentissage, une base de données d'images omnidirectionnelles pour la détection de personnes. En effet, les bases de données conventionnelles ne sont pas adaptées à la géométrie des caméras omnidirectionnelles. La base de données d'images omnidirectionnelles a été générée à partir de la base de données perspective. En effet, chaque image perspective de personne a été utilisée pour générer une série d'images omnidirectionnelles. Lorsque la détection est réalisée dans l'espace de la sphère unitaire, une seule image sphérique est nécessaire puisque l'image sphérique est invariante à la rotation. Les résultats obtenus ont montré que le descripteur sphérique obtient de meilleurs taux de détection comparé au descripteur omnidirectionnel.

Nos travaux ont pour objectif la valorisation du capteur omnidirectionnel dans cette tâche de suivi 3D. Ce type de capteur avec son large champ de vision peut d'une part limiter la sortie de champ d'une personne et d'autre part, avec les adaptions nécessaires de son modèle de formation, présenter des résultats comparables à une utilisation de caméra perspective.

Nous avons ensuite proposé une approche 3D de détection et de suivi de personnes. Elle est basée sur l'implémentation de fonctions de vraisemblances adaptées à l'espace de la sphère, permettant un suivi 3D dans le cas des images omnidirectionnelles. L'approche

suivi de 3D que nous avons proposée est basée sur un modèle de personne à 32 degrés de liberté. Plusieurs fonctions de vraisemblance ont été considérées. Les premières sont basées sur les distances géodésiques dans l'espace sphérique $\mathcal{S}^2$. Les secondes sont basées sur la mise en correspondance de la silhouette dans l'image sphérique et le modèle 3D projeté sur la sphère unitaire. Les fonctions de vraisemblance combinées à un filtre particulaire ont permis d'obtenir un suivi 3D précis de la personne.

Limitations et perspectives :

Le système que nous avons présenté dans le cadre du suivi 3D présente quelques limitations :

1- La contrainte mono-capteur, ne nous permet pas d'avoir d'information de profondeur. Nous avons été amenés à instaurer des contraintes fortes pour la taille de notre modèle 3D. Cette dernière est supposée être connue ou bien obtenue grâce à une connaissance a priori de la taille d'une personne moyenne et du rapport des différents membres.

2- Notre système ne traite pas le cas des occultations totales. En effet si la personne à suivre est occultée, le système de suivi sera fortement perturbé. Cela surtout si la personne réapparait dans une zone où elle n'était pas présente avant l'occultation (par exemple, la personne passe derrière un objet statique occultant).

3- Notre base de test n'a pas exploré de geste complexe dans un contexte de changement de luminosité comme cela peut être le cas dans le cas de déplacement libre dans un espace extérieur.

Quelques perspectives :

Pour les futurs travaux, il serait intéressant d'explorer les méthodes de modèle déformable

lors de la détection. Cela permettrait dans notre cas de faire du suivi 3D basé sur un apprentissage. Ainsi, les travaux futurs pourront inclure un terme de compatibilité basé sur l'apprentissage pour gérer plus efficacement les occultations et améliorer la précision avec laquelle la profondeur des membres est estimée. Cette phase d'apprentissage nous permettra d'intégrer dans le suivi 3D un système de décision de type réseaux de neurones pour faire une correspondance entre la détection 2D et 3D. Car nous pensons qu'il peut être intéressant d'extraire davantage d'information de la détection 2D qu'uniquement la position globale de la personne dans une zone. Ainsi, l'utilisation d'un modèle déformable pouvant détecter la position 2D de quelque articulations du corps pourrait contribuer de manière plus significative au suivi 3D. Toujours dans le le dessein d'améliorer le système de suivi 3D dans le cas des déplacements complexes de la personne à suivre, il est nécessaire d'enrichir la base de données 3D avec des mouvements plus complexes lors du calcul de la matrice de covariance de la dynamique de mouvement du sujet.

Dans le cas de la détection 2D, il serait intéressant de comparer la détection avec un apprentissage fait à partir d'une base de données d'apprentissage directement collectée avec un capteur catadioptrique. Par ailleurs, nous avons tout au long de la thèse concentré nos efforts uniquement sur les descripteurs basés sur le gradient. Il sera intéressant de mettre en place d'autres descripteurs tels que le LBP présenté dans le cas des images perspectives au cas des capteurs catadioptriques. Cela sera parfaitement possible étant donné que l'on peut aisément mettre en place des méthodes permettant de déterminer le voisinage d'un point sur la sphère.

Notre objectif à court terme est d'intégrer les algorithmes et méthodes développés dans des applications plus complexes, comme la reconnaissance de geste dans un contexte d'interaction Homme-Machine. La difficulté réside dans le temps de calcul qui demeure trop élevé pour un traitement temps réel. Pour répondre à ces contraintes de calcul, il sera

nécessaire de changer de langage de programmation en évitant l'utilisation de Matlab pour certaines fonctions. De plus il serait intéressant de mettre en place une implémentation basée sur les cartes graphiques à l'aide d'outil tel que CUDA. De plus, les travaux futurs se concentreront sur l'amélioration de l'interpolation réalisée lors de la transition de l'image omnidirectionnelle à l'image sphérique.

Annexe A

Calibrage du système Smarttrack - caméra omnidirectionnelle

Dans cette partie nous allons montrer comment on a effectué le calibrage de la caméra omnidirectionnelle avec le Smarttrack. Dans un premier temps on suppose que la caméra omnidirectionnelle est calibrée et que les paramètres intrinsèques sont connues. De même le système Smarttrack propriétaire est calibré. Les points 3D sont obtenus par rapport à un repère monde défini lors de son calibrage. L'objectif ici est de faire correspondre les points 3D obtenus par le Smarttrack au points correspondant dans l'image omnidirectionnelle. Ce qui revient à déterminer la matrice de passage [R,t] du repère monde choisi lors du calibrage du Smarttrack au repère caméra. Cela peut être vu comme la recherche des paramètres extrinsèques de notre caméra omnidirectionnelle avec un repère mire qui correspond au repère monde du Smarttrack. Pour cela on va extraire de manière manuelle m point 3D p_i et leurs valeurs e_i qui correspondent à leurs projections dans l'image. La fonction à

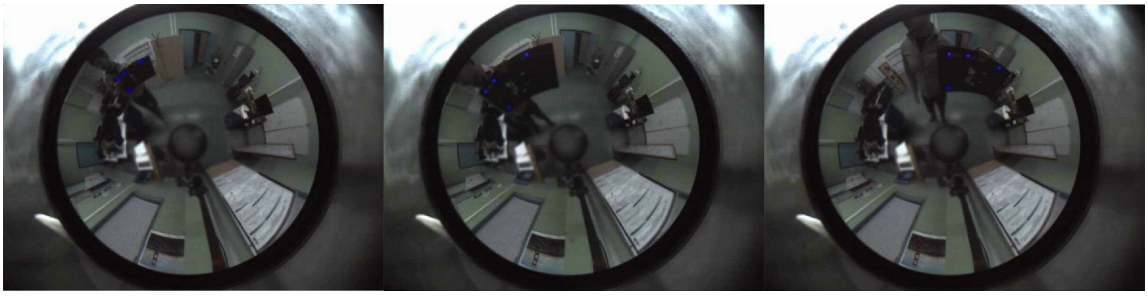


FIGURE A.1 – Projection des marqueurs 3D sur l'image omnidirectionnelle dans une séquence de test composée de 30 images. .

minimiser pour retrouver la matrice de passage est :

$$F(x) = 1/2 \sum_{i=1}^{m} \|e_i - P(K, R, t, p_i)\|^2 \tag{A.1}$$

Avec K la matrice des paramètres intrinsèques de la caméra . R, t sont la rotation et la translation de la pose de la caméra utilisée pour définir le paramètres extrinsèques par rapport au repère monde du capteur Smarttrack. P est la projection du point du modèle p dans l'image de la caméra. Pour cela nous avons utilisé l'algorithme de Levenberg-Marquardt pour l'optimisation de cette fonction de coût. l'algorithme de Levenberg-Marquardt, est une descente de gradient basée sur la méthode des moindres carrés. Cette étape de minimisation non-linéaire ajuste le calibrage de notre capteur plus précisément et permet une meilleure mise en correspondance entre la caméra et le Smarttrack, ce qui résulte en une meilleure correspondance entre les points 3D et 2D. Comme on peut le voir sur la figure A.1 cela permet de déterminer avec précision les paramètres $[R, t]$ permettant de rendre l'erreur de projection négligeable ainsi elle n'influence pas les résultats obtenus par notre algorithme de suivi 3D. Au début nous avions supposé que les paramètres de calibrage intrinsèques sont connues, cela n'est pas nécessaire. Seule une estimation de ces paramètres est nécessaire car lors de la minimisation de la fonction de coût on peut améliorer la

précision des paramètres intrinsèques également.

Bibliographie

[Abd-Almageed et al., 2007] Abd-Almageed, W., Hussein, M., and Abdelkader, M. (2007). Real-time human detection and tracking from mobile vehicles. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, pages 149–154. IEEE.

[Alonso et al., 2007] Alonso, I. P., Llorca, D. F., Sotelo, M. Á., Bergasa, L. M., de Toro, P. R., Nuevo, J., Ocaña, M., and Garrido, M. Á. G. (2007). Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2) :292–307.

[Ardouin, 2014] Ardouin, J. (2014). *Contribution à l'étude de la visualisation d'environnements réels et virtuels avec un champ visuel étendu.* PhD thesis, INSA de Rennes.

[Arulampalam et al., 2002] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2) :174–188.

[Azad et al., 2004] Azad, P., Ude, A., Dillmann, R., and Cheng, G. (2004). A full body human motion capture system using particle filtering and on-the-fly edge detection. In *Humanoid Robots, 2004 4th IEEE/RAS International Conference on*, volume 2, pages 941–959. IEEE.

[Baker and Nayar, 1998] Baker, S. and Nayar, S. K. (1998). A theory of catadioptric image formation. In *Computer Vision, 1998. Sixth International Conference on*, pages 35–42. IEEE.

[Baker and Nayar, 1999] Baker, S. and Nayar, S. K. (1999). A theory of single-viewpoint catadioptric image formation. *International Journal of Computer Vision*, 35(2) :175–196.

[Balan et al., 2005] Balan, A. O., Sigal, L., and Black, M. J. (2005). A quantitative evaluation of video-based 3d person tracking. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 349–356. IEEE.

[Barreto and Araujo, 2001] Barreto, J. P. and Araujo, H. (2001). Issues on the geometry of central catadioptric image formation. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE.

[Barreto et al., 2003] Barreto, J. P., Martin, F., and Horaud, R. (2003). Visual servoing/tracking using central catadioptric images. In *Experimental Robotics VIII*, pages 245–254. Springer.

[Barrow et al., 1977] Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., and Wolf, H. C. (1977). Parametric correspondence and chamfer matching : Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER.

[Belongie and Malik, 2000] Belongie, S. and Malik, J. (2000). Matching with shape contexts. In *2000 Proceedings Workshop on Content-based Access of Image and Video Libraries*, pages 20–26.

[Belongie et al., 2002] Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4) :509–522.

[Bertozzi et al., 2007] Bertozzi, M., Broggi, A., Del Rose, M., Felisa, M., Rakotomamonjy, A., and Suard, F. (2007). A pedestrian detector using histograms of oriented gradients and a support vector machine classifier. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, pages 143–148. IEEE.

[Bo and Sminchisescu, 2010] Bo, L. and Sminchisescu, C. (2010). Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1) :28–52.

[Bogdanova et al., 2007] Bogdanova, I., Bresson, X., Thiran, J.-P., and Vandergheynst, P. (2007). Scale space analysis and active contours for omnidirectional images. *IEEE Transactions on Image Processing*, 16(7) :1888–1901.

[Bregler et al., 2004] Bregler, C., Malik, J., and Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3) :179–194.

[Carnegie-Mellon,] Carnegie-Mellon, U. Cmu graphics lab motion capture database. [Online ; accessed 2017-11-18].

[Caron et al., 2009] Caron, G., Marchand, E., and Mouaddib, E. M. (2009). 3d model based pose estimation for omnidirectional stereovision. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 5228–5233. IEEE.

[Chua et al., 2015] Chua, J.-L., Chang, Y. C., and Lim, W. K. (2015). A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing*, 9(3) :623–633.

[Cielniak et al., 2005] Cielniak, G., Treptow, A., and Duckett, T. (2005). Quantitative performance evaluation of a people tracking system on a mobile robot. In *Proc. 2nd European Conference on Mobile Robots*.

[Coelingh et al., 2010] Coelingh, E., Eidehall, A., and Bengtsson, M. (2010). Collision warning with full auto brake and pedestrian detection-a practical example of automatic emergency braking. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 155–160. IEEE.

[Conseil, 2008] Conseil, S. (2008). *Suivi tridimensionnel de la main et reconnaissance de gestes pour les Interfaces Homme Machine*. PhD thesis, Université Paul Cézanne-Aix-Marseille III.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3) :273–297.

[Corvee and Bremond, 2010] Corvee, E. and Bremond, F. (2010). Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 469–475. IEEE.

[Dai et al., 2005] Dai, C., Zheng, Y., and Li, X. (2005). Layered representation for pedestrian detection and tracking in infrared imagery. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 13–13. IEEE.

[Dalal, 2005] Dalal, N. (2005). INRIA Person Dataset. [Online ; accessed 19-July-2017].

[Dalal, 2006] Dalal, N. (2006). *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

[Delibasis et al., 2016] Delibasis, K. K., Georgakopoulos, S. V., Kottari, K., Plagianakos, V. P., and Maglogiannis, I. (2016). Geodesically-corrected zernike descriptors for pose recognition in omni-directional images. *Integrated Computer-Aided Engineering*, 23(2) :185–199.

[Deutscher et al., 2000] Deutscher, J., Blake, A., and Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 126–133. IEEE.

[Deutscher and Reid, 2005] Deutscher, J. and Reid, I. (2005). Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2) :185–205.

[Duchoň et al., 2012] Duchoň, F., Hubinskỳ, P., Hanzel, J., Babinec, A., and Tölgyessy, M. (2012). Intelligent vehicles as the robotic applications. *Procedia Engineering*, 48 :105–114.

[Felzenszwalb et al., 2010a] Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2010a). Cascade object detection with deformable part models. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 2241–2248. IEEE.

[Felzenszwalb et al., 2010b] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010b). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9) :1627–1645.

[Felzenszwalb and Huttenlocher, 2005] Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International journal of computer vision*,

61(1) :55–79.

[Fiala and Basu, 2003] Fiala, M. and Basu, A. (2003). Robot navigation using panoramic landmark tracking. *TECHNICAL PAPERS-SOCIETY OF MANUFACTURING ENGINEERS-ALL SERIES-*.

[Fiala and Basu, 2005] Fiala, M. and Basu, A. (2005). Panoramic stereo reconstruction using non-svp optics. *Computer Vision and Image Understanding*, 98(3) :363–397.

[Fontmarty, 2008] Fontmarty, M. (2008). *Vision et filtrage particulaire pour le suivi tridimensionnel de mouvements humains : applications à la robotique*. PhD thesis, Université Paul Sabatier-Toulouse III.

[Freeman et al., 1991] Freeman, W. T., Adelson, E. H., et al. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9) :891–906.

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

[Georgakopoulos et al., 2016] Georgakopoulos, S. V., Kottari, K., Delibasis, K., Plagianakos, V. P., and Maglogiannis, I. (2016). Convolutional neural networks for pose recognition in binary omni-directional images. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 106–116. Springer.

[Gerónimo et al., 2007] Gerónimo, D., López, A., Ponsa, D., and Sappa, A. (2007). Haar wavelets and edge orientation histograms for on–board pedestrian detection. *Pattern Recognition and Image Analysis*, pages 418–425.

[Geyer and Daniilidis, 2000] Geyer, C. and Daniilidis, K. (2000). A unifying theory for central panoramic systems and practical implications. *Computer Vision—ECCV 2000*, pages 445–461.

[González et al., 2001] González, R. C., Valdés, R., and Cancelas, J. A. (2001). Vision based measurement system to quantify straightness defect in steel sheets. In *International Conference on Computer Analysis of Images and Patterns*, pages 427–434. Springer.

[He and Wang, 1990] He, D.-C. and Wang, L. (1990). Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, 28(4) :509–512.

[Holte et al., 2012] Holte, M. B., Tran, C., Trivedi, M. M., and Moeslund, T. B. (2012). Human pose estimation and activity recognition from multi-view videos : Comparative explorations of recent developments. *IEEE Journal of selected topics in signal processing*, 6(5) :538–552.

[Isard and Blake, 1998] Isard, M. and Blake, A. (1998). Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1) :5–28.

[Jain et al., 2000] Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition : A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1) :4–37.

[Jeong et al., 2004] Jeong, C.-Y., Kim, J.-S., and Hong, K.-S. (2004). Appearance-based nude image detection. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 467–470. IEEE.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1) :35–45.

[Kang et al., 2002] Kang, S., Byun, H., and Lee, S.-W. (2002). Real-time pedestrian detection using support vector machines. *Pattern Recognition with Support Vector Machines*, pages 273–281.

[Lan and Huttenlocher, 2005] Lan, X. and Huttenlocher, D. P. (2005). Beyond trees : Common-factor models for 2d human pose recovery. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 470–477. IEEE.

[Laptev, 2006] Laptev, I. (2006). Improvements of object detection using boosted histograms. In *BMVC*, volume 3, pages 949–958.

[Lee and Elgammal, 2010] Lee, C.-S. and Elgammal, A. (2010). Coupled visual and kinematic manifold models for tracking. *International Journal of Computer Vision*, 87(1) :118–139.

[Lee and Cohen, 2004] Lee, M. W. and Cohen, I. (2004). Proposal maps driven mcmc for estimating human body pose in static images. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE.

[Leibe et al., 2005] Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 878–885. IEEE.

[Levi and Weiss, 2004] Levi, K. and Weiss, Y. (2004). Learning object detection from a small number of examples : the importance of good features. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE.

[Lienhart and Maydt, 2002] Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002*

International Conference on, volume 1, pages I–I. IEEE.

[Lindman et al., 2010] Lindman, M., Ödblom, A., Bergvall, E., Eidehall, A., Svanberg, B., and Lukaszewicz, T. (2010). Benefit estimation model for pedestrian auto brake functionality. In *Proceedings of the 4th International Conference Expert Symposium on Accident Research*.

[Mariano et al., 2002] Mariano, V. Y., Min, J., Park, J.-H., Kasturi, R., Mihalcik, D., Li, H., Doermann, D., and Drayer, T. (2002). Performance evaluation of object detection algorithms. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 965–969. IEEE.

[Mei, 2007] Mei, C. (2007). *Couplage Vision Omnidirectionnelle et Télémétrie Laser pour la Navigation en Robotique/Laser-Augmented Omnidirectional Vision for 3D Localisation and Mapping*. PhD thesis, INRIA Sophia Antipolis, Project-team ARobAS.

[Mei and Rives, 2007] Mei, C. and Rives, P. (2007). Single view point omnidirectional camera calibration from planar grids. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 3945–3950. IEEE.

[Menezes et al., 2006] Menezes, P., Lerasle, F., and Dias, J. (2006). Visual tracking modalities for a companion robot. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5363–5368. IEEE.

[Micušık, 2004] Micušık, B. (2004). *Two-view geometry of omnidirectional cameras*. PhD thesis, PhD. Thesis, Czech Technical University.

[Mikolajczyk et al., 2004] Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. *Computer Vision-ECCV 2004*, pages 69–82.

[Mori et al., 2005] Mori, G., Belongie, S., and Malik, J. (2005). Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11) :1832–1837.

[Mu et al., 2008] Mu, Y., Yan, S., Liu, Y., Huang, T., and Zhou, B. (2008). Discriminative local binary patterns for human detection in personal album. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

[Nalwa, 1996] Nalwa, V. (1996). A true omnidirectional viewer. Technical report, technical report, Bell Laboratories.

[Navaratnam et al., 2007] Navaratnam, R., Fitzgibbon, A. W., and Cipolla, R. (2007). The joint manifold model for semi-supervised multi-valued regression. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

[Navaratnam et al., 2005] Navaratnam, R., Thayananthan, A., Torr, P. H., and Cipolla, R. (2005). Hierarchical part-based human body pose estimation. In *BMVC*.

[Nene and Nayar, 1998] Nene, S. A. and Nayar, S. K. (1998). Stereo with mirrors. In *Computer Vision, 1998. Sixth International Conference on*, pages 1087–1094. IEEE.

[Nie et al., 2017] Nie, B. X., Wei, P., and Zhu, S.-C. (2017). Monocular 3d human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision*.

[Ning et al., 2008] Ning, H., Xu, W., Gong, Y., and Huang, T. (2008). Discriminative learning of visual words for 3d human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

[Ojala et al., 1996] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59.

[Ojala et al., 2002] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7) :971–987.

[Oren et al., 1997] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio, T. (1997). Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 193–199. IEEE.

[Pajdla et al., 2001] Pajdla, T., Svoboda, T., and Hlaváč, V. (2001). Epipolar geometry of central panoramic catadioptric cameras. In *Panoramic vision*, pages 73–102. Springer.

[Papageorgiou and Poggio, 2000] Papageorgiou, C. and Poggio, T. (2000). A trainable system for object detection. *International Journal of Computer Vision*, 38(1) :15–33.

[Papageorgiou et al., 1998] Papageorgiou, C. P., Oren, M., and Poggio, T. (1998). A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE.

[Paul et al., 2013] Paul, M., Haque, S. M., and Chakraborty, S. (2013). Human detection in surveillance videos and its applications-a review. *EURASIP Journal on Advances in Signal Processing*, 2013(1) :176.

[Peleg et al., 2001a] Peleg, S., Ben-Ezra, M., and Pritch, Y. (2001a). Omnistereo : Panoramic stereo imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3) :279–290.

[Peleg et al., 2001b] Peleg, S., Ben-Ezra, M., and Pritch, Y. (2001b). Panoramic imaging with horizontal stereo. In *Panoramic vision*, pages 143–160. Springer.

[Peursum et al., 2010] Peursum, P., Venkatesh, S., and West, G. (2010). A study on smoothing for particle-filtered 3d human body tracking. *International Journal of Computer Vision*, 87(1) :53–74.

[Poppe, 2010] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6) :976–990.

[Porikli, 2005] Porikli, F. (2005). Integral histogram : A fast way to extract histograms in cartesian spaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 829–836. IEEE.

[Puig et al., 2014] Puig, L., Guerrero, J. J., and Daniilidis, K. (2014). Scale space for camera invariant features. *IEEE Transactions on pattern analysis and machine intelligence*, 36(9) :1832–1846.

[Scaramuzza et al., 2006a] Scaramuzza, D., Martinelli, A., and Siegwart, R. (2006a). A flexible technique for accurate omnidirectional camera calibration and structure from motion. In *Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on*, pages 45–45. IEEE.

[Scaramuzza et al., 2006b] Scaramuzza, D., Martinelli, A., and Siegwart, R. (2006b). A toolbox for easily calibrating omnidirectional cameras. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5695–5701. IEEE.

[Scaramuzza and Siegwart, 2007] Scaramuzza, D. and Siegwart, R. Y. (2007). A new method and toolbox for easily calibrating omnidirectional cameras.

[Shashua et al., 2004] Shashua, A., Gdalyahu, Y., and Hayun, G. (2004). Pedestrian detection for driving assistance systems : Single-frame classification and system level performance. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 1–6. IEEE.

[Shet et al., 2007] Shet, V. D., Neumann, J., Ramesh, V., and Davis, L. S. (2007). Bilattice-based logical reasoning for human detection. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

[Sidenbladh, 2004] Sidenbladh, H. (2004). Detecting human motion with support vector machines. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 188–191. IEEE.

[Sigal et al., 2010] Sigal, L., Balan, A. O., and Black, M. J. (2010). Humaneva : Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1) :4–27.

[Sigal et al., 2004] Sigal, L., Bhatia, S., Roth, S., Black, M. J., and Isard, M. (2004). Tracking loose-limbed people. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[Sminchisescu and Telea, 2002] Sminchisescu, C. and Telea, A. (2002). Human pose estimation from silhouettes. a consistent approach using distance level sets. In *10th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG'02)*, volume 10.

[Sminchisescu and Triggs, 2003a] Sminchisescu, C. and Triggs, B. (2003a). Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6) :371–391.

[Sminchisescu and Triggs, 2003b] Sminchisescu, C. and Triggs, B. (2003b). Kinematic jump processes for monocular 3d human tracking. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[Srinivasan and Shi, 2007] Srinivasan, P. and Shi, J. (2007). Bottom-up recognition and parsing of the human body. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 153–168. Springer.

[Taiana et al., 2010] Taiana, M., Santos, J., Gaspar, J., Nascimento, J., Bernardino, A., and Lima, P. (2010). Tracking objects with generic calibrated sensors : An algorithm based on color and 3d shape features. *Robotics and autonomous systems*, 58(6) :784–795.

[Vesanto and Alhoniemi, 2000] Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3) :586–600.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[Viola and Jones, 2002] Viola, P. and Jones, M. (2002). Fast and robust classification using asymmetric adaboost and a detector cascade. In *Advances in neural information processing systems*, pages 1311–1318.

[Wachter and Nagel, 1997] Wachter, S. and Nagel, H.-H. (1997). Tracking of persons in monocular image sequences. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 2–9. IEEE.

[Wang and He, 1990] Wang, L. and He, D.-C. (1990). Texture classification using texture spectrum. *Pattern Recognition*, 23(8) :905–910.

[Yagi and Kawato, 1990] Yagi, Y. and Kawato, S. (1990). Panorama scene analysis with conic projection. In *Intelligent Robots and Systems' 90.'Towards a New Frontier of Applications', Proceedings. IROS'90. IEEE International Workshop on*, pages 181–187. IEEE.

[Yang and Ramanan, 2013] Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12) :2878–2890.

[Yasin et al., 2016] Yasin, H., Iqbal, U., Kruger, B., Weber, A., and Gall, J. (2016). A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956.

[Ying and Hu, 2004] Ying, X. and Hu, Z. (2004). Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model. *Computer Vision-ECCV 2004*, pages 442–455.

[Yu et al., 2017] Yu, J., Sun, J., Liu, S., and Luo, S. (2017). Multi-activity 3d human motion recognition and tracking in composite motion model with synthesized transition bridges. *Multimedia Tools and Applications*, pages 1–33.

[Zaklouta and Stanciulescu, 2012] Zaklouta, F. and Stanciulescu, B. (2012). Real-time traffic-sign recognition using tree classifiers. *IEEE Transactions on Intelligent Transportation Systems*, 13(4) :1507–1514.

[Zhang and Fan, 2010] Zhang, X. and Fan, G. (2010). Dual gait generative models for human motion estimation from a single camera. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4) :1034–1049.

[Zhou et al., 2012] Zhou, S., Liu, Q., Guo, J., and Jiang, Y. (2012). Roi-hog and lbp based human detection via shape part-templates matching. In *Neural Information Processing*, pages 109–115. Springer.

[Zhu et al., 2009] Zhu, J., Zou, H., Rosset, S., Hastie, T., et al. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3) :349–360.

[Zhu et al., 2006] Zhu, Q., Yeh, M.-C., Cheng, K.-T., and Avidan, S. (2006). Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1491–1498. IEEE.