Intégration de méthodes de représentation et de classification pour la détection et la reconnaissance d'obstacles dans des scènes routières

Bassem Besbes

Institut National des Sciences Appliquées de Rouen

Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes

Thèse de Doctorat

Discipline : INFORMATIQUE

Présentée par

Bassem Besbes

Pour obtenir le titre de docteur de l'INSA de Rouen

Intégration de méthodes de représentation et de classification pour la détection et la reconnaissance d'obstacles dans des scènes routières

Soutenue le 16/09/2011 devant le jury composé de :

Jacques Jacot	Laboratoire de Production Microtechnique, EPFL, Suisse	Président
Fabrice Meriaudeau	Laboratoire Le2i, Université de Bourgogne	Rapporteur
Fawzi Nashashibi	Centre de recherche INRIA, Paris-Rocquencourt	Rapporteur
Pierre Bonton	Laboratoire LASMEA, Université Blaise Pascal	Examinateur
Abdelaziz Bensrhair	Laboratoire LITIS, INSA de Rouen	Directeur de thèse
Alexandrina Rogozan	Laboratoire LITIS, INSA de Rouen	Encadrant
Julien Rebut	VALEO	Invité

Remerciements

C'est avec bonheur que je consacre ces mots en signe de reconnaissance à tous ceux qui ont contribué, de prés ou de loin, à la réalisation de cette thèse. Qu'ils veuillent apercevoir ici mes termes les plus sincères de remerciements.

Mes premiers remerciements vont vers mon directeur de thèse Monsieur Abdelaziz Bensrhair. Je voudrais le remercier pour l'écoute, la confiance, la générosité, la bonne humeur qu'il m'a accordées tout au long de ces années. Il m'a souvent aidé à surmonter les difficultés et m'a fourni d'excellentes conditions logistiques et financières. Je remercie bien entendu Madame Alexandrina Rogozan, qui en agissant à titre d'encadrant a su m'initier à la recherche et à me pousser à toujours faire mieux.

Je remercie chaleureusement Messieurs Fabrice Meriaudeau et Fawzi Nashashibi pour avoir accepté le difficile rôle de rapporteur de ces travaux ainsi que Pierre Bonton, Jacques Jacot et Julien Rebut d'avoir accepté de prendre part à mon jury.

Je remercie Monsieur Yousri Kessentini pour sa collaboration dans le co-encadrement du stage de Sonda Ammar. Je remercie également les stagiaires avec qui j'ai travaillé : Abir Zribi, Sonda Ammar et Amine Azzaoui.

Je tiens à remercier le directeur du laboratoire LITIS Monsieur Stéphane Canu, pour son accueil et son écoute. Je souhaite aussi remercier l'ensemble des enseignants du département ASI de l'INSA de Rouen pour leur soutien et leurs conseils avisés tout au long de trois ans de monitorat, particulièrement Monsieur Nicolas Delestre, Alexandre Pauchet et Thierry Le Pors.

Merci au personnel administratif et technique du LITIS, particulièrement Brigitte Diarra, Sandra Hagues et Jean François Brulard pour leur disponibilité et

leur bonne humeur tout au long de ces années.

Une pensée émue pour mes amis du laboratoire LITIS que j'ai été amené à côtoyer. Un merci plus particulier à Florian Yger, Rémi Flamary, Benjamin Labbé, Yacine Sid Ahmed, Amnir Hadachi et Carlo Abi Chahine pour leurs relectures, leurs précieux conseils et avec qui les échanges furent toujours chaleureux et constructifs.

Un grand merci aux familles Chtiwi, Ayadi et Kessentini qui m'ont encouragé et soutenu tout au long des années que j'ai passées en France, sachez que vos encouragements n'ont pas été vains, et ont contribué à cet aboutissement.

Je voudrais aussi témoigner ma gratitude envers mes enseignants, ainsi que toutes les personnes qui ont contribué à ma formation en Tunisie et en France.

Je tiens à remercier tous les membres de ma famille pour m'avoir soutenu depuis 27 ans et qui ne cessent de m'encourager à aller plus loin. Un grand merci à ma tante Tahia pour sa relecture et ses corrections. Enfin, ma reconnaissance la plus profonde s'adresse à mes parents pour leurs sacrifices, leurs soutiens inconditionnels et leur appui moral permanent, malgré la distance qui me sépare d'eux.

Table des matières

Table des figures

Glossaire

VIS	Visible
IR	Infrarouge
SWIR	Infrarouge proche
LWIR	Infrarouge lointain
OR	les obstacles routiers
SURF	Speeded Up Robust Features
SVM	Machines à vecteurs de support (Support Vector Machine)
POI	Point d'Intérêt
ROI	Région d'Intérêt (Region of Interest)
DST	Théorie de Dempster-Shafer
MCT	Modèle des Croyances Transférables
VV	Vocabulaire Visuel
VVH	Vocabulaire Visuel Hiérarchique
LMK	Noyaux par mise en correspondance (Local Matchnig Kernel)

Intégration de méthodes de représentation et de classification pour la détection et la reconnaissance d'obstacles dans des scènes routières

Résumé Cette thèse s'inscrit dans le contexte de la vision embarquée pour la détection et la reconnaissance d'obstacles routiers, en vue d'application d'assistance à la conduite automobile.

À l'issue d'une étude bibliographique, nous avons constaté que la problématique de détection d'obstacles routiers, notamment des piétons, à l'aide d'une caméra embarquée, ne peut être résolue convenablement sans recourir aux techniques de reconnaissance de catégories d'objets dans les images. Ainsi, une étude complète du processus de la reconnaissance est réalisée, couvrant les techniques de représentation, de classification et de fusion d'informations. Les contributions de cette thèse se déclinent principalement autour de ces trois axes.

Notre première contribution concerne la conception d'un modèle d'apparence locale basé sur un ensemble de descripteurs locaux SURF (Speeded Up Robust Features) représentés dans un Vocabulaire Visuel Hiérarchique. Bien que ce modèle soit robuste aux larges variations d'apparences et de formes intra-classe, il nécessite d'être couplé à une technique de classification permettant de discriminer et de catégoriser précisément les objets routiers. Une deuxième contribution présentée dans la thèse porte sur la combinaison du Vocabulaire Visuel Hiérarchique avec un classifieur SVM.

Notre troisième contribution concerne l'étude de l'apport d'un module de fusion multimodale permettant d'envisager la combinaison des images visibles et infrarouges. Cette étude met en évidence de façon expérimentale la complémentarité des caractéristiques locales et globales ainsi que la modalité visible et celle infrarouge. Pour réduire la complexité du système, une stratégie de classification à deux niveaux de décision a été proposée. Cette stratégie est basée sur la théorie des fonctions de croyance et permet d'accélérer grandement le temps de prise de décision.

Une dernière contribution est une synthèse des précédentes : nous mettons à profit les résultats d'expérimentations et nous intégrons les éléments développés dans un système de détection et de suivi de piétons en infrarouge-lointain. Ce système a été validé sur différentes bases d'images et séquences routières en milieu urbain.

Mots clés : Vision embarquée, Détection et reconnaissance d'obstacles routiers, Représentation des images, Classification par SVM, Fusion de capteurs, Fonction de croyances, Détection de piétons en Infrarouge-lointain.

Integrating Representation and Classification Methods for Obstacle detection in road scenes

Abstract The aim of this thesis arises in the context of Embedded-vision system for road obstacles detection and recognition : application to driver assistance systems.

Following a literature review, we found that the problem of road obstacle detection, especially pedestrians, by using an on-board camera, cannot be adequately resolved without resorting to object recognition techniques. Thus, a preliminary study of the recognition process is presented, including the techniques of image representation, Classification and information fusion. The contributions of this thesis are organized around these three axes. Our first contribution is the design of a local appearance model based on SURF (Speeded Up Robust Features) features and represented in a hierarchical Codebook. This model shows considerable robustness with respect to significant intra-class variation of object appearance and shape. However, the price for this robustness typically is that it tends to produce a significant number of false positives. This proves the need for integration of discriminative techniques in order to accurately categorize road objects. A second contribution presented in this thesis focuses on the combination of the Hierarchical Codebook with an SVM classifier.

Our third contribution concerns the study of the implementation of a multimodal fusion module that combines information from visible and infrared spectrum. This study highlights and verifies experimentally the complementarities between the proposed local and global features, on the one hand, and visible and infrared spectrum on the other hand. In order to reduce the complexity of the overall system, a two-level classification strategy is proposed. This strategy, based on belief functions, enables to speed up the classification process without compromising the recognition performance. A final contribution provides a synthesis across the previous ones and involves the implementation of a fast pedestrian detection system using a far-infrared camera. This system was validated with different urban road scenes that are recorded from an onboard camera.

Key words : Embedded vision, Road obstacle detection and recognition, Image representation, SVM classification, Sensor fusion, Belief functions, Pedestrian detection in far-infrared images.

Chapitre 1

Introduction et positionnement de la thèse

Introduction

Ce premier chapitre introductif décrit les motivations scientifiques et le positionnement de cette thèse. Au cours de ce chapitre, nous commençons par décrire le contexte et les enjeux liés aux systèmes de vision pour la détection et la reconnaissance des obstacles routiers. Dans un deuxième temps, nous essayons d'identifier les problématiques et d'analyser les solutions proposées en littérature. Enfin, nous justifions notre choix méthodologique et nous présentons les bases d'image sur lesquelles nous avons mené les expérimentations.

1.1 Contexte

Avec l'augmentation constante du trafic routier, le risque d'accidents augmente également. Toutes les statistiques de la sécurité routière montrent que presque 10 millions de personnes dans le monde sont chaque année impliquées dans un accident de la route. Ces accidents causent plus de 1,2 millions de personnes. Les chiffres sont énormes et alarmants. Les premières actions des constructeurs automobiles ont porté sur la réduction des conséquences de collisions. Ces actions se sont concrétisées par l'intégration de dispositifs de sécurité comme les ceintures de sécurité, les attaches et ancrages inférieurs pour les sièges d'enfants et les coussins gonflables (airbags). Néanmoins, ces dispositifs ne permettent que de réduire la gravité des accidents. Depuis 1990, on a pu observer d'autres mesures plus avancées comme les freins antiblocages (ABS), les systèmes de traction asservie, les systèmes de surveillance de la pression des pneus, etc. Ces systèmes sécuritaires

participent à la réduction de la proportion et de la gravité des accidents puisqu'ils interviennent en amont de l'accident.
Durant la dernière décennie, la recherche s'est penchée sur des systèmes non seulement sécuritaires mais aussi intelligents. Des constructeurs automobiles, des équipementiers et des laboratoires de recherche se sont rassemblés autour du concept de *Véhicule Intelligent*. On dit *Intelligent* parce que le développement du véhicule repose sur des fonctions généralement associées à l'intelligence : capacités sensorielles, mémoire et utilisation des technologies récentes de l'information et de la communication. Les projets déjà montés dans ce domaine sont nombreux, nous en citons une liste non exhaustive : Arco (2001), eSAfety(2006), e-MOTION (2003), MobiVIP (2005), LOVe (2006). D'autres recherches focalisées sur l'analyse des causes des accidents ont montré que l'inattention, le manque de vigilance et la défaillance du jugement du conducteur sont les principales sources d'accidents. C'est à l'examen de ces points que se révèle l'importance des systèmes intelligents d'aide à la conduite. Composés de systèmes visant à assister le conducteur sur différentes dimensions de la conduite, les systèmes intelligents d'aide à la conduite présentent un enjeu majeur du point de vue sécurité routière.

1.1.1 Les systèmes intelligents d'aide à la conduite

La voiture de demain sera intelligente et le conducteur pourra bénéficier d'une assistance accrue d'aide à la conduite. Cette assistance est fondée sur le développement de systèmes embarqués capables de fournir en temps réel des informations utiles au conducteur afin de faciliter sa tâche, d'optimiser sa prise de décision et de sécuriser ses déplacements. La plupart des véhicules actuels disposent de systèmes de freinage antiblocage, d'alerte de vitesse, de surveillance de la pression des pneus et d'autres. Ces systèmes permettent de fournir une assistance de bas niveau vu qu'ils n'utilisent que des informations inhérentes au véhicule. Pour une assistance de plus haut niveau, les informations liées à l'environnement proche du véhicule comme les bords de route, les obstacles routiers, la distance d'éloignement sont indispensables afin d'assister le conducteur notamment dans des situations difficiles (détection d'obstacles, vision nocturne, régulation de vitesse, etc). Pour ces raisons les constructeurs d'automobile sont de plus en plus demandeurs de systèmes de haut niveau et surtout intelligents. Les systèmes actuels de haut niveau sont connus sous l'acronyme *ADAS* (Advanced Driver Assistance Systems). Ils assistent le conducteur dans sa prise de décision, lui transmettent un signal d'alerte en cas de situation dangereuse et peuvent même exécuter des actions afin

d'éviter l'accident. Parmi ces systèmes nous pouvons citer les deux fameux systèmes l'ACC (Adaptative Cruise Control) et le LDW (Lane Departure Warning). Le premier permet de maintenir une distance de sécurité entre les véhicules en adaptant automatiquement la vitesse du véhicule. Quant au deuxième, il permet le maintien de la voie de circulation et réagit lorsque le conducteur sort de son couloir par inadvertance. Le progrès technologique est incontestable, mais il est si incontestable que les voitures de demain intègrent des modules de détection des obstacles routiers notamment des usagers vulnérables comme les piétons et les cyclistes. Ces modules sont d'absolue nécessité pour éviter les collisions et pour sauver ainsi des vies.

1.1.2 Intégration de modules de détection d'obstacles routiers

Dans le cadre de la conception de systèmes d'aide à la conduite automobile, l'intégration d'un module de détection d'obstacles est une tâche essentielle. Ce module pourra aider le conducteur dans sa perception, car malheureusement, il n'est pas toujours vigilant. Les facteurs qui peuvent intervenir dans la perte de vigilance du conducteur sont la fatigue occasionnée par une conduite de nuit, ou une conduite prolongée. De plus, plusieurs événements perturbateurs peuvent déconcentrer le conducteur, lorsqu'il règle le son de l'autoradio par exemple, ou quand il parle au téléphone ou à un autre passager. Plusieurs situations dangereuses pourraient être évitées si le conducteur reçoit d'avance une alerte. Une petite histoire pourrait en résumer l'importance : Un conducteur est pressé d'arriver à un meeting où il est attendu pour donner un discours important à des clients. Il s'arrête avec hésitation au feu rouge qui vient juste de s'allumer. Il n'est pas sûr de savoir comment mieux présenter son produit. Il est très important que ses clients soient intéressés, il pourrait conclure une grosse vente. Peut-être devrait-il présenter son produit comme étant leur unique solution, ou bien leur meilleur choix. Le feu passe au vert, il accélère, tourne en trombe au coin de la rue, espérant que les clients vont tolérer son retard. Sortant à toute vitesse du virage derrière la maison, il ne voit pas la fille qui traverse la rue en courant. Quelques minutes plus tard, l'ambulance arrive... Une situation pareille est malheureusement très fréquente. En France, 12% des causes de décès d'accidents routiers sont des piétons renversés par des voitures.
L'intégration d'un module intelligent de détection vise à éviter les collisions avec n'importe quel type d'obstacle (voiture, piéton, cycliste, animal, ...). La perception de l'environnement routier est assurée par l'implantation d'un système de

capteurs embarqué jouant le rôle des organes de sens chez l'homme. Le traitement des données issues des capteurs par des algorithmes temps-réel permet de détecter et d'identifier de manière fiable les obstacles routiers (On notera désormais OR obstacles routiers). Le système pourra intervenir dans des situations à risques en utilisant différents déclencheurs tels que l'assistance au freinage ou le freinage autonome. Ainsi, l'intervention du système permet d'éviter la collision ou d'atténuer considérablement l'impact en réduisant la vitesse du véhicule avant la collision. Si l'accident ne peut être évité, des déclencheurs de protection peuvent être activés (comme les airbags). Toutes ces mesures permettront d'assurer une conduite plus sécurisée et de sauver ainsi des vies.
Bien que dans la littérature nous trouvons une très grande variété de systèmes de détection d'obstacles, il n'existe jusqu'à nos jours, aucun système qui a pu être commercialisé. Pourtant dans le cadre des projets de recherches, de nombreux capteurs actifs et passifs ont été utilisés pour percevoir l'environnement. Les méthodes proposées pour traiter les données afin de détecter les obstacles sont également très diverses. Dans la section suivante, nous faisons le point sur les capteurs embarqués et sur les méthodes de détection spécifiques à leur traitement.

1.1.2.1 Les capteurs embarqués

Dans ce paragraphe, nous donnons une vue d'ensemble des principaux capteurs utilisés pour la détection d'OR. Les capteurs peuvent être regroupés, d'une part, en capteurs proprioceptifs/extéroceptifs. D'autre part, ils peuvent être aussi classés comme étant actifs ou bien passifs. Les capteurs proprioceptifs sont capables de mesurer un attribut en fonction de leur propre état. Cet attribut peut être l'accélération, l'orientation ou la vitesse de l'objet sur lequel ils sont montés. Par exemple, les capteurs d'inclinaison, les accéléromètres et les odomètres [1] sont des capteurs proprioceptifs. Par contre, les capteurs extéroceptifs sont capables de mesurer un attribut d'un objet externe présent dans la scène. Les caméras qui fonctionnent dans le spectre visible ou infrarouge, les sonars, les Lidars et les Radars sont des exemples de capteurs extéroceptifs. Deux approches prédominent dans la perception de l'environnement d'un véhicule par un capteur extéroceptif : les capteurs actifs et la vision par capteurs passifs. Le capteur actif transmet un signal dans un environnement donné et mesure ensuite l'interaction de ce signal avec l'environnement (exemple : Lidar, Radar, sonar). À la différence des capteurs actifs, les capteurs passifs récupèrent l'information de manière non-intrusive. Ainsi, ils

1. Des instruments qui mesurent la distance parcourue

n'émettent pas de radiation, ils ne font donc que recevoir un signal qui peut être réfléchi, émis ou transmis par des sources d'énergie externes. Les caméras visibles et certaines caméras infrarouges sont des exemples de capteurs passifs. Dans la suite, nous dressons les points forts et les points faibles des principaux capteurs utilisés dans les systèmes de détection d'OR avant de justifier le choix des capteurs que nous allons utiliser.

Les Radars

Le Radar est un système composé principalement d'une antenne émettrice/réceptrice d'une onde. Il émet des ondes radios ou des radiations microondes en une série de pulsions à partir de l'antenne et reçoit la partie d'énergie qui est réfléchie par la cible. Le temps nécessaire à l'onde pour voyager de la cible jusqu'à l'objet permet de déterminer la distance et la vitesse (dans le cas de plusieurs émissions) de l'objet.
Le Radar a sa propre source d'énergie qui peut pénétrer à travers les nuages et la pluie. Ainsi, il est considéré comme un détecteur toute saison. De plus, il a une portée très élevée offrant la possibilité de détecter des objets très distants. Tous ces avantages ont contribué de façon évidente à l'apparition de systèmes ACC avec Radar en option dans des véhicules haut de gamme (par exemple Mercedes classe S).
Néanmoins, la faible résolution spatiale (surtout dans le sens latéral) du Radar entraîne des détections moins fiables voir même inexistantes pour les petits obstacles. En outre, les parties métalliques présentent entre autres une réflectivité supérieure comparée aux autres objets comme les êtres humains. Ainsi, les objets qui présentent une forte réflexion minimisent l'effet des autres réflexions moins fortes et conduisent donc à des fausses détections. Enfin, un inconvénient majeur du Radar est le problème d'interférences qui se dégrade en présence de plusieurs voitures utilisant la même technologie dans le trafic.

Les Lidars

Le principe de fonctionnement du Lidar (appelé aussi télémètre laser) est basé sur la mesure du temps mis par la lumière réfléchie sur l'obstacle se trouvant dans l'axe de tir du laser. Les Lidars se basent sur le même principe que les Radars étant donné que la distance aux objets est déterminée par le temps séparant les pulsions transmises et reçues. Ils sont utilisés pour des distances d'environ 40m et ont une grande précision dans les deux directions : longitudinale et latérale. Généralement,

les méthodes utilisées pour détecter les obstacles sont similaires à celles utilisées en traitement d'images : la segmentation, le clustering et le tracking. Toutefois, le contenu des images Lidar est différent de celui des images visibles. En effet, le Lidar fournit une image de profondeur, tandis que les caméras captent la réflexion de la lumière visible.
Ce capteur est largement utilisé par la communauté robotique pour la navigation de robots en terrain inconnu. Néanmoins, le coût, l'encombrement et la consommation d'énergie élevée limitent son utilisation en embarqué sur un véhicule.

Les caméras visibles

Une caméra reçoit des énergies émises sans qu'elle même n'irradie la scène. Les images capturées par les caméras visibles (On notera désormais VIS la modalité visible), en couleurs ou en niveaux de gris, sont d'un côté très riches en contenu, de l'autre, difficiles à interpréter. Ceci est peut être la raison pour laquelle la recherche s'est beaucoup focalisée dans cette direction. Dans la littérature, on trouve une très grande variété d'approches, de techniques et d'algorithmes de traitement d'images qui ont été proposés pour la détection, le suivi et la reconnaissance d'objets dans les images. De plus, avec les avancées technologiques, les caméras visibles sont devenues moins chères et faciles à embarquer sur des véhicules.
Deux approches sont possibles en perception passive : l'utilisation d'une seule caméra visible ou l'exploitation de plusieurs points de vue avec plusieurs caméras reliées de manière rigide sur le véhicule. La vision monoculaire consiste à équiper le véhicule avec une seule caméra présentant des avantages de coût et de simplicité de mise en œuvre. Néanmoins, l'inconvénient de cette méthode est qu'elle ne permet pas de restituer la profondeur de la scène observée. L'utilisation de deux caméras (technique de stéréovision) permet, quant à elle, d'accéder à l'information tridimensionnelle.
Le principe de la stéréovision est d'inférer de l'information sur la structure et les distances 3D d'une scène à partir de deux images optiques prises de points de vue différents. La stéréovision dont nous exposerons le principe se déroule en trois étapes successives : calibrage, appariement et triangulation. La mise en correspondance entre les images gauches et droites (appariement) est la phase du traitement la plus difficile. En effet, les deux images de la scène peuvent présenter de grandes différences aussi bien en terme de morphologie que d'illumination puisque les caméras sont décalées. Un point localisé dans l'une des images peut se retrouver sans homologue dans l'autre image à cause des problèmes de recouvrement ou d'occlu-

sion. En outre, il est difficile de repérer des indices visuels permettant d'effectuer la mise en correspondance dans une scène faiblement texturée. Ces circonstances expliquent pourquoi aucun système basé sur la stéréovision n'a pu être commercialisé.
Bien que les caméras visibles aient été plus étudiées par rapport aux caméras infrarouges, les caméras sensibles au spectre visible souffrent de limitations liées aux conditions climatiques et aux conditions d'illumination (surtout pendant la nuit). Ces difficultés peuvent être surmontées grâce à l'utilisation de caméras infrarouges.

Les caméras infrarouges

Les caméras infrarouges ont été utilisées dans une grande variété d'applications. Le spectre infrarouge (On notera désormais IR la modalité infrarouge) est typiquement subdivisé en bandes, dont la séparation n'est pas bien définie et varie selon les auteurs. Dans le domaine de détection d'obstacles routiers, deux technologies ont été considérées pour des applications liées principalement à la détection de piétons. Il s'agit de l'infrarouge actif et l'imagerie thermique, qui correspondent aux bandes réflectives et thermiques. La différence principale entre l'imagerie en bandes réflectives et thermales est que la première retient les informations réfléchies par les objets, quant à la seconde, elle enregistre la température émise par les objets.
L'utilisation de l'infrarouge actif est limitée car ses performances dépendent de plusieurs facteurs comme les changements des conditions d'illumination, de forme, de vitesse et de la couleur de l'objet à détecter. Ce sont les caméras thermiques qui sont fortement utilisées avec des longueurs d'ondes longues offrant une perception large de l'environnement routier. Ces caméras sont appelées aussi capteurs infrarouges passifs, puisqu'ils capturent les rayonnements infrarouges émises par les objets chauds sans utiliser une source artificielle d'illumination.
La loi de Planck [2] (Max Planck 1858-1947) montre que les distributions de l'énergie selon la longueur d'onde se retrouvent toujours sous une même forme et que pour chaque longueur d'onde la luminance augmente avec la température. Les caméras thermiques sont en fait des capteurs permettant la mesure d'une luminance. Ils permettent de transformer une image captée dans le domaine infrarouge et fonction de la luminance de l'objet observé, en une image visible et analysable par l'oeil humain. L'avantage majeur des caméras thermiques est qu'elles peuvent produire des images lisibles dans l'obscurité complète. Leur portée est dépendante

2. http ://fr.wikipedia.org/wiki/Loi_de_Planck

des conditions atmosphériques, du type de la caméra et de la différence de température de la cible avec le fond. Le brouillard et la pluie limitent cette portée car le rayonnement IR peut être affaibli. Mais toutefois, cette portée reste plus grande dans la bande IR que dans le VIS.
On distingue entre deux types de caméras thermiques infrarouges : refroidies et non-refroidies. Les caméras thermiques refroidies ont un capteur intégré à un système de refroidissement qui fait descendre la température du capteur à une valeur très basse. Ainsi, le bruit issu de la chaleur de fonctionnement du capteur demeure inférieur au signal qui provient de la scène observée. Ces caméras peuvent être utilisées pour produire des images dans les infrarouges moyens (MWIR, longueur d'onde entre $3-5\mu$ m). D'une manière générale, les images nocturnes d'une caméra MWIR sont plus contrastées que celles produites dans une autre bande de l'infrarouge.
Les caméras infrarouges non-refroidis comportent souvent un micro bolomètre : une minuscule résistance qui fait que toute variation de température dans la scène observée provoque une variation de la température du bolomètre. Cette variation est convertie en un signal électrique, qui est utilisé pour améliorer l'image. Les capteurs non refroidis sont conçus pour travailler dans l'infrarouge lointain (LWIR, longueur d'onde entre $7-14\mu$ m) où les objets terrestres émettent la plus grande part de leur énergie infrarouge. La radiation thermique qui émane des êtres humains et des animaux est à son maximum entre 8 et 14μ m. Ainsi, ces objets présentent un contraste plus important dans les images produites dans la bande LWIR.
Il existe une troisième bande d'IR à courte longueur d'onde (entre $1-2\mu$ m) appelée SWIR. Cette bande se propage mieux au travers des atmosphères humides et sera donc choisi de préférence pour des applications maritimes. Par temps de pluie, la portée de la caméra est à peu près la même en LWIR et SWIR. Par ailleurs, le LWIR traverse mieux les fumées et sa portée en brouillard est plus grande.

La figure 1.1 présente différentes prises d'une image avec des longueurs d'ondes différentes et montre que le contraste change en fonction des longueurs d'ondes. Il est indéniable que plus le contraste thermique est élevé, plus il est facile de détecter des cibles sur un fond de température constante. C'est ainsi que la grande majorité des systèmes qui sont apparus sur des véhicules, ces dernières années, reposent sur l'utilisation de caméras infrarouges de type LWIR. Tout récemment, des constructeurs automobiles comme Honda, Mercedes et BMW commercialisent un assistant de vision nocturne. Ce dernier permet de restituer sur l'écran central

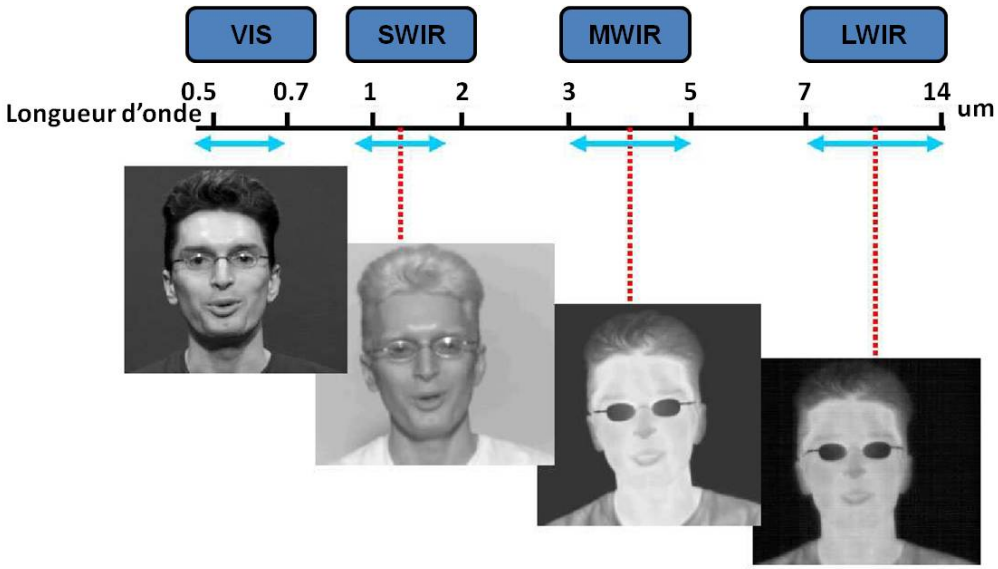


Figure 1.1. Présentation de l'image d'un visage prise avec différentes longueurs d'ondes

une image routière où les piétons et les animaux présentent les zones les plus claires de l'image.
Bien que l'avantage de l'utilisation des caméras infrarouges pour la détection de piétons soit évident, les véhicules et d'autres obstacles n'émettant pas de chaleur, sont difficilement repérables dans ces images. En tout cas, chaque capteur peut s'apercevoir d'informations invisibles à d'autres capteurs. C'est ce qui fait l'intérêt de la fusion de capteurs que nous détaillons le principe dans la section suivante.

1.1.2.2 Fusion de capteurs

La fusion concerne l'utilisation de différentes informations provenant de différents capteurs pour obtenir une image meilleure de l'environnement routier. Dans un tel système dédié à la détection d'obstacles routiers (détection, reconnaissance, suivi), la collaboration de différents capteurs permet d'en accroître ses performances.
Deux capteurs différents peuvent détecter le même objet avec une précision différente par rapport aux paramètres qui décrivent l'objet. Ces informations sont complémentaires et permettent une meilleure mesure au niveau de l'intégrité et de la précision. De plus, certains capteurs peuvent s'apercevoir d'informations invisibles à d'autres capteurs comme l'information de la profondeur des objets qui est fournie directement par un capteur actif. Les principales difficultés liés à la fusion

de capteurs concernent le calibrage automatique et le prototypage des algorithmes de fusion.
Les systèmes fondés sur la fusion de capteurs sont généralement robustes mais très chers. À notre connaissance, le seul système commercialisé est développé par l'entreprise Mobileye [3]. Ce système a été équipé pour les marques de Volvo s60, lancées en Avril 2010. Ce système permet d'activer un freinage d'urgence lorsque le conducteur ne réagit pas à temps devant un piéton détecté par le système constitué d'un Lidar et d'une camera visible. Cependant, les risques des interférences liées à l'utilisation simultanée de capteurs actifs ne sont pas écartés. Mais, les principaux problèmes concernent le calibrage des capteurs et le prototypage des algorithmes de fusion. C'est ainsi que nous avons été amenés à opter pour l'utilisation de capteurs de vision.

1.1.2.3 Choix de capteur de vision

Le choix du (ou des) capteur(s) qu'il convient d'utiliser pour percevoir l'environnement est une étape cruciale dans la réalisation d'un système de détection des OR. La sélection doit prendre en compte le domaine de fonctionnement de chaque capteur et de ses performances. Nous commençons d'abord par comparer les deux catégories : les capteurs actifs et les capteurs passifs. Les capteurs actifs fonctionnent même dans les conditions climatiques dégradées, ou dans de mauvaises conditions d'illumination comme la nuit. Ils fournissent directement la profondeur des obstacles. Cette information utile est très pertinente pour l'étape de détection. Néanmoins, les capteurs actifs ne peuvent pas détecter des objets de petites tailles représentés par un nombre réduit de points. Dans le cas de la détection de piétons, un objet de petite taille peut être un enfant, chose qui ne serait pas tolérable. De plus, les capteurs actifs ne sont pas très adaptés pour l'étape de reconnaissance car l'information récupérée est difficile à analyser. En effet, elle est beaucoup moins riche comparée à celle récupérée par une caméra. Rajoutons à tout cela l'inconvénient majeur du coût de ces capteurs ainsi que le problème des interférences qui peuvent apparaître si plusieurs véhicules en sont équipés. Par ailleurs, les capteurs passifs sont moins onéreux et les systèmes basés sur la vision ont démontré leur efficacité pour des applications d'aide à la conduite.
Le choix d'un système en monovision semble logique si on envisage d'implémenter une technique à la fois rapide et moins chère. De plus, l'utilisation d'une configuration intégrant deux ou plusieurs capteurs entraîne des difficultés techniques

3. http ://en.wikipedia.org/wiki/Mobileye

importantes. Nous allons à présent comparer les deux capteurs passifs : les caméras visibles et les caméras infrarouges. Les caméras visibles sont moins onéreuses que les caméras infrarouges. Cependant, elles souffrent des mêmes limitations que la visibilité humaine dans les milieux dégradés (conditions de forte pluie, de brouillard ou durant la nuit). Ceci n'est pas le cas des caméras infrarouges qui peuvent continuer à faire de la détection même dans des conditions climatiques difficiles et pendant la nuit. Elles sont certes légèrement plus chères que les caméras visibles, mais leur prix est en baisse grâce aux avancées technologiques.
Toutes ces considérations nous conduisent à opter pour un système monovision utilisant une caméra infrarouge. Afin de justifier ce choix, nous présentons également dans ce manuscrit des expérimentations qui ont été faites non seulement sur des images en IR, mais aussi sur un système combinant les deux caméras. La problématique de l'interprétation systématique des images d'un environnement routier est présentée dans la section suivante.

1.2 Vision pour la détection et la reconnaissance des obstacles routiers

La détection d'obstacles par un système de monovision embarqué sur un véhicule est une problématique difficile. En effet, le trafic routier implique un nombre variable d'objets différents. Les arbres, le mobilier urbain, les piétons, les cyclistes et les véhicules sont tous des objets pouvant être présents dans une scène routière. Nous désignons par obstacles routiers les objets qui se situent sur la trajectoire d'un véhicule qui sont majoritairement des véhicules et des piétons. La figure 1.2 montre deux images, en IR et en VIS, extraites d'une scène routière.

Figure 1.2. Exemple d'images VIS et IR extraites d'une scène routière

Durant les quinze dernières années, de nombreux travaux de référence ont été effectués aussi bien sur la détection d'obstacles routiers que sur leur classification et leur reconnaissance. Malgré les progrès techniques considérables dans le domaine de la vision par ordinateur, les problématiques ne sont pas encore résolues, et du coup, aucun système n'a pu être industrialisé. Selon la façon dont les systèmes échouent, les obstacles non détectés exposent le conducteur à un risque sérieux d'accident. Dans la section suivante, nous précisons les exigences que doit satisfaire un système de détection et de reconnaissance d'obstacles routiers.

1.2.1 Exigences d'un système de détection d'obstacles routiers

Dans ce paragraphe, nous donnons les principales exigences auxquelles un système de détection et de reconnaissance d'obstacles doit répondre pour pouvoir être considéré comme une solution fiable. Un tel système embarqué, en vue d'aider le conducteur dans sa tâche de conduite, doit répondre aux exigences suivantes :

- La robustesse,

Le système doit être capable de détecter des obstacles quelque soient leurs apparences, échelles et formes. Il doit aussi être robuste aux différentes conditions d'illumination et répondre aux problèmes d'occultation qui accentuent autant la variabilité des objets routiers. Le piéton est l'objet le plus difficile à détecter en raison notamment de la grande variabilité d'apparences et de l'articulation du corps humain. Un piéton peut se présenter avec différentes tenues vestimentaires, tenant des accessoires différents (parapluie, chapeau,...) et dans des scènes complexes notamment avec les phénomènes de foule. Bien que ces difficultés soient moins marquées pour les voitures, la grande variabilité de types de voitures (voitures de tourisme, bus, camions, tracteurs, ...) rend leur détection et leur reconnaissance difficiles.

Toutes ces difficultés s'intensifient en présence de fond encombré ou dans les conditions métrologiques dégradées produisant du bruit ou un manque de contrastante dans les images.

- L'efficacité et la précision,

Le système doit détecter les objets sur la route de manière fiable et précise. En effet, les obstacles non détectés exposent le conducteur à un risque sérieux d'accident. Quant aux fausses alertes, elles poussent le conducteur à ne plus avoir confiance dans le système. Ainsi, un tel système doit détecter tous les obstacles sans commettre aucune erreur, quelles que soient les conditions environnementales et la configuration de la scène routière. Concernant l'étape de la reconnaissance,

le système doit identifier rapidement la classe d'appartenance de l'obstacle afin de pouvoir donner au conducteur une marge de manœuvre adéquate.

- La contrainte du temps réel,

Le terme temps réel possède plusieurs significations suivant le contexte. Dans notre contexte de travail, nous considérons qu'un système est temps réel si l'information après son traitement reste pertinente. En d'autres termes, le système permet d'avertir le conducteur avant que ce dernier puisse réagir. Généralement, la cadence de traitement de 10 images est compatible avec les exigences temps réel (délai de réponse proche de 100 ms). C'est avec de tels systèmes que l'on peut s'autoriser aujourd'hui l'implantation de systèmes embarqués à bord de véhicules permettant de fournir des fonctions d'aide à la conduite

- Le coût,

Un système embarqué sur un véhicule doit être beaucoup moins chère que le prix du véhicule. Notre système utilise une seule caméra infrarouge, qui certes coûte légèrement plus cher qu'une caméra visible, mais dont le prix reste bien moins élevé que celui de capteurs actifs comme le Radar ou le Lidar.

1.2.2 Etat de l'art sur les méthodes de détection d'obstacles routiers

Cette section n'a pas pour but de constituer un état de l'art exhaustif sur les différentes techniques permettant de détecter les obstacles routiers dans les images. Toutefois, nous cherchons à faire un tour d'horizon des principales méthodes utilisées pour les systèmes monovision, notamment en infrarouge. La figure 1.3 donne un exemple d'une scène routière filmée en IR où les obstacles sont correctement détectés.

La majorité des systèmes développés jusqu'à présent reposent sur trois étapes : la génération d'hypothèses, leur validation et leur suivi. La première étape consiste à localiser les endroits qui contiennent éventuellement des obstacles. Nous les appellerons dorénavant des régions d'intérêt (ROI : Regions Of Interest). Dans cette étape, l'algorithme bien qu'il analyse tous les pixels de l'image, doit rester efficace en terme de temps de calcul. Ceci est problématique car les obstacles, notamment les piétons, peuvent se retrouver dans n'importe quel endroit dans l'image. Généralement, les méthodes employées procèdent à segmenter l'image totale selon un critère défini a apriori. En effet, la segmentation est une étape essentielle en traitement d'images dans la mesure où elle conditionne l'interprétation de régions spécifiques dans ces images.

Figure 1.3. Exemple d'une scène routière en IR (bande LWIR) où les obstacles sont correctement détectés. Les piétons et les voitures sont respectivement encadrés par des rectangles (fenêtres englobantes) en rouge et en bleu.

Le problème principal qui découle de l'utilisation d'un seul capteur de vision est qu'aucune information de profondeur n'est fournie. Les systèmes basés sur l'utilisation de capteurs actifs (Lidar ou Radar) à balayage ou de la stéréovision permettent de générer une carte de profondeur en 3D. Ainsi, le critère de définition des régions d'intérêt correspond à la position des pixels dans le monde réel. Les obstacles sont détachés du fond de l'image en utilisant l'approche région. L'avantage de cette solution réside dans l'extraction des informations pertinentes tout en éliminant les caractéristiques appartenant à la scène elle-même. Cette approche comporte toutefois des limites associées plus particulièrement avec les objets complexes qui se recouvrent.

Généralement, la définition de ROI avec un système monovision est basée sur la segmentation mouvement ou la recherche de primitives ou d'indices spécifiques du type d'obstacle à détecter.

Les techniques de segmentation du mouvement [ELW03, DPKA04] sont généralement employées pour des caméras fixes où deux images successives sont prises dans les mêmes conditions d'acquisition. De plus, seuls les objets en mouvement peuvent être détectés et un piéton ou un véhicule immobile dans une zone de collision ne seront pas détectés. Cette limitation liée à la position de la caméra ne nous permet donc pas d'envisager une application embarquée.

La recherche de primitives caractéristiques du type de l'obstacle à détecter consiste

à rechercher des formes, des apparences ou des indices particuliers. Comme par exemple, faire apparaitre les objets symétriques, identifier les structures rectangulaires afin de détecter les véhicules, ou mettre en évidence les régions de fortes intensités, dont le but de localiser des régions de piétons dans les images infrarouges.
Le développement dans le domaine d'apprentissage assisté par l'ordinateur a permis d'aborder d'autres techniques de détection de catégories d'objets dans les images. Nous en citons d'une façon particulière les nouvelles méthodes basées sur l'apprentissage d'un modèle implicite de formes ou ceux basés sur l'utilisation de classifieurs. Ces techniques font recours à un apprentissage hors ligne permettant d'identifier les primitives caractéristiques des objets en question.
Généralement, les méthodes de détection qui n'utilisent pas de classifieurs ne sont pas suffisantes pour éliminer les fausses alertes. Ainsi, l'usage d'autres processus est souvent nécessaire : la validation par des techniques de vérification (vérification de certaines propriétés, corrélation avec des modèles, ...) ou par des méthodes de classification, ou suivi des hypothèses de détection permettent de confirmer la présence d'obstacles et de rejeter les fausses détections. Dans la suite nous détaillons les différentes méthodes de définition des régions d'intérêt, de validation d'hypothèses et de suivi.

1.2.2.1 Génération de ROI

Il existe plusieurs méthodes de génération de ROI, la plus simple est de sélectionner toutes les ROI de l'image. Cela nécessite de balayer la totalité de l'image dans toutes les directions et à plusieurs valeurs d'échelles. Ceci rend le traitement par la suite, très coûteux en termes de nombre d'opérations de calcul. Dans cette section, nous présentons les méthodes les plus fréquemment citées dans la littérature.

Les primitives et les indices caractéristiques d'obstacles à détecter

Il existe différentes méthodes permettant d'extraire des éléments caractéristiques des obstacles routiers dans les images. Sans être exhaustif, nous pouvons citer différents travaux qui peuvent s'inscrire dans ce thème.
Les images de véhicules et piétons sont en général symétriques horizontalement et verticalement. Cette observation a été utilisée comme indice dans de nombreux travaux pour la détection de piéton et de véhicule [BBFL02, BBFV06, TBM$^+$06]. Dans [BBFN00], les régions d'intérêt sont définies en analysant une carte de sy-

métrie construite en examinant les critères de symétrie des niveaux de gris et des contours de l'image. Ces techniques sont généralement utilisées pour définir des ROI correspondantes à des véhicules. Du fait des tailles réduites de piétons dans les images, l'analyse de symétrie est utilisée surtout pour la vérification de leurs présences [BBF+04].
En plus du critère de symétrie, autres travaux ont cherché à identifier des formes rectangulaires pour la détection de véhicules. Dans ce sens, Bertke et al [BHD00] proposent un algorithme qui permet de détecter les véhicules lointains. L'algorithme proposé cherche à identifier une structure rectangulaire en se basant sur le nombre de points de contours horizontaux et verticaux. D'autres travaux se sont focalisés sur la recherche des indices portées d'ombres (extraction par seuillage des niveaux de gris) [TS98] ou de feux arrières (extraction à partir de la composante rouge de l'image couleur) [CTB05] pour détecter les véhicules. Néanmoins, ces derniers indices sont particulièrement utiles pour le traitement des scènes d'autoroute en VIS. Dans la section suivante, nous présentons les techniques appropriées au traitement des images en IR.

Détection d'obstacles par fenêtre glissante

La technique la plus générique est la détection par fenêtre glissante. Cela consiste à parcourir exhaustivement l'image en appliquant un détecteur à de très nombreuses positions et échelles. Cette technique est appliquée surtout en l'absence de toutes informations permettant de localiser des endroits contenant des obstacles. La méthode propose donc de balayer l'image et de comparer le contenu de chaque fenêtre avec un modèle spécifique d'un obstacle routier. Parmi les travaux les plus marquants, nous pouvons citer les travaux de Gavrilla [Gav00, GGM04] qui reposent sur l'extraction de contours et la corrélation avec des modèles de piétons. D'une façon générale, les méthodes basées sur la mise en correspondance avec un modèle explicite utilisent des seuils empiriques qui peuvent être sources de plusieurs problèmes. Le réglage empirique d'un tel paramètre n'est pas précis et nécessite un certain temps, en fonction du nombre de modèles utilisés et du taux d'acceptation d'erreur.
Contrairement aux techniques précédemment citées, les méthodes basées sur la classification constituent eux-mêmes, par apprentissage, les modèles qu'ils utilisent dans leur recherche. Ces modèles se basent sur l'extraction d'un jeu de caractéristiques pour représenter un tel objet. Les caractéristiques les plus connues à cet égard permettent de caractériser la forme et la texture des objets [OPS+97, DT05].

Les algorithmes de classification sont basés avant tout sur des méthodes d'analyse statistique et sont restreints à déterminer l'appartenance ou non d'une image à une classe. Ainsi, utilisé de manière isolée, un tel algorithme de classification d'une ROI est incapable de localiser les obstacles. Toutefois, le parcours exhaustif de l'image à l'ensemble des dimensions et des positions (fenêtres) possibles avec des classifieurs le permet. Généralement, les méthodes utilisant des classifieurs sont très robustes mais l'exploration de l'image entière induit des temps de calcul très importants. Dans la pratique, le recours à des méthodes de classification rapide comme l'utilisation de SVM linéaire [Vap95] et les techniques à plusieurs cascades de classifieurs boostés [JVJS03, NCHP08] permettent d'accélérer le processus de détection. En outre, la limitation de la région de recherche en présence d'informations sur les voies de circulation [BLS09] ou sur la perspective de l'image permet d'envisager des applications rapides, voire en temps réel.

Apprentissage d'un modèle implicite de formes

Les techniques que nous avons présentées dans les sections précédentes se basent soit sur la recherche de primitives caractéristiques, soit sur la comparaison avec des modèles d'analyse et de représentation de forme ou de texture. Pour certains objets, comme le piéton, il est impossible de définir un modèle explicite qui permet de gérer à la fois les grandes variations intra-classe et les problèmes d'occultations prononcées. Afin de répondre à ces problématiques, Leibe et al [LLS04, Lei08] proposent une approche basée sur l'apprentissage d'un modèle de forme implicite. Ce modèle est établi par apprentissage et peut être défini comme un modèle de distribution non-paramétrique d'un ensemble de motifs qui caractérisent l'apparence locale de piétons (voir figure 1.4). Ces motifs peuvent être représentés par des voisinages locaux (des patches) [LLS04] ou des descripteurs extraits autour des points d'intérêt [MLS06].

Lors de la phase de détection, la mise en correspondance des motifs détectés dans une image avec le modèle implicite de forme construit par apprentissage, permet d'une part de déterminer des probabilités pour chaque motif appartenant ou pas à un piéton ; d'autre part, de voter pour la position des centres des objets. Ces votes sont ensuite interprétés dans un cadre probabiliste en utilisant la transformée de Hough généralisée et l'algorithme de Mean Shift [Com03].
Cette approche est très intéressante dans le sens où la méthode de détection est positionnée de façon à résoudre le problème d'occultation. De plus, cette approche se distingue clairement des méthodes classiques car elle ne requiert ni une étape

Figure 1.4. Un exemple de modèle d'apparences regroupant un ensemble de motifs qui caractérisent les apparences locales de piétons

préalable de segmentation ni un parcours exhaustif de l'image pour chercher les piétons. En effet, le traitement principal se focalise sur la détection et l'appariement des motifs extraits autour de points d'intérêt, robustes à l'échelle et la rotation. Enfin, notons que cette approche a été utilisée avec succès dans des applications récentes de détection de piétons non seulement en VIS [LSS05] mais aussi en IR [JA09].

Techniques appropriées au traitement des images en IR

L'image produite par une caméra infrarouge est basée sur l'émission de chaleur dans la scène, apportant ainsi des informations différentes de celles du spectre visible. L'utilisation de ce capteur permet ainsi d'acquérir visuellement les sources de chaleur émises par les objets et les corps présents dans la scène. Cette propriété est très utile en pratique pour détecter dans les images infrarouges des zones particulières, significatives de la présence de piétons.
Plusieurs techniques de détection de piétons en infrarouge commencent par faire un seuillage pour ne garder que les pixels ayant des valeurs d'intensité élevée. Une valeur de seuil doit être assez élevée pour éliminer les pixels de l'arrière-plan et assez basse pour laisser apparaître la majorité des pixels représentant le piéton. Les manières de déterminer les seuils sont différents, mais la majorité se base sur le calcul de seuils adaptatifs [XLF]. Ensuite, les sources de chaleur sont localisées

en cherchant les colonnes d'image contenant le plus de pixels clairs.
Notons que les méthodes dans le domaine visible peuvent aussi s'appliquer : utilisation d'un modèle implicite de formes [JA09], de classifieurs [SRBB06] ou la recherche d'indices caractéristiques. Dans [MR04], les auteurs montrent qu'en procédant par la détection de têtes, les piétons occultés ou présents dans des situations de foules peuvent être détectés séparément. Ainsi, l'indice caractéristique recherché correspond à une région claire ayant une forme circulaire dans l'image.

1.2.2.2 Vérification et validation des hypothèses de détection

La validation des hypothèses de détection est une étape cruciale avant toute mesure de prise de décision faisant l'objet d'alerte ou de contrôle de véhicule. Cette étape permet non seulement de vérifier la présence d'un obstacle routier mais aussi de déterminer sa classe d'appartenance : piéton, véhicule ou mauvaise détection.
Après l'étape de détection, des régions d'intérêt sont définies contenant d'éventuels obstacles routiers. Quoi qu'il en soit, il est difficile d'affirmer l'existence de telle catégorie d'objet routier contenu dans la ROI. Il est vrai que l'utilisation de l'IR permet d'acquérir visuellement des sources de chaleur émises par les objets et les corps présents dans la scène. En revanche, des objets comme les animaux, les véhicules, les motos, les boîtes électriques, des lampadaires, ... pourrait produire des zones de lumière supplémentaire dans les images infrarouges. Les techniques classiques de vérification de certaines propriétés (dimensions, ratio, symétrie) [BBFL02, XLF] ou la mise en correspondance avec un modèle-objet [BBF+04, GGM04] ne sont pas assez robustes pour résoudre un problème de discrimination multiclasses. En outre, ils ne permettent pas de faire face à une large variation de formes et d'apparences intra-classe. Afin de surmonter ces problèmes, les techniques d'apprentissage automatique sont actuellement de plus en plus appliquées à cet égard [DGL07, YJZ10]. Ces techniques sont fondées sur l'utilisation d'algorithmes qui s'adaptent à identifier des apparences, des formes, des textures caractéristiques d'un tel objet, grâce à une phase initiale d'apprentissage.

1.2.2.3 Suivi des obstacles routiers

Une scène routière filmée par une caméra peut être vue comme une succession d'images. Si la fréquence d'acquisition est élevée, les déplacements des objets entre deux images successives sont faibles. Le suivi d'objet permet non seulement d'estimer ces déplacements mais aussi d'affiner le processus de détection.
L'estimation du déplacement et la localisation instantanée sont des informations

utiles qui permettent de souligner la trajectoire d'un objet en mouvement et de prédire ses éventuels déplacements. Le suivi de trajectoires se fait généralement par l'application de filtre temporel comme le filtre de Kalman [Kal60] ou à particule [GSS93]. Ces filtres ont deux phases distinctes :

- la prédiction : permet de produire une estimation de l'état courant,
- la mise à jour : permet de corriger l'état prédit dans le but d'obtenir une estimation plus précise.

Ils utilisent ainsi un vecteur d'état sur chaque nouvelle image, à partir d'une initialisation définie manuellement ou automatiquement.
Dans le cadre de suivi d'obstacles routiers, le processus de suivi est implémenté plutôt afin de raffiner les détections et d'éviter la lourde tâche de re-détection d'objets dans chaque image. En effet, ce processus consomme moins de ressources par rapport à la détection. De plus, il peut prendre le relais au cas où la détection échoue temporairement, par exemple à cause d'un problème d'occultation.
Dans ce cas, il suffit d'appliquer des techniques de mise en correspondance qui ne nécessitent pas forcément une étape de mise à jour. En effet, le processus de mise en correspondance peut être omis non seulement pour des considérations de temps de calcul, mais aussi car la durée d'observation d'objets vulnérables est généralement très courte. L'algorithme le plus connu à cet égard est celui de Mean Shift [FH75] qui utilise une procédure itérative de recherche de maximum local dans un espace $\mathbb{R}^d$. Cet algorithme nécessite une initialisation manuelle de zone de recherche et prend pour entrée une carte représentative de la distribution de caractéristiques. Dans [CRM00, Com03], l'algorithme de Mean Shift a été utilisé pour le suivi d'objets déformables. Dans chaque image, l'algorithme recherche la position de l'objet à partir des mesures de similarité entre des distributions de couleurs.
Dans un tel algorithme de suivi, il convient de mettre au point un ensemble de caractéristiques pertinentes représentatives de l'objet. Ce qui revient à dire, qu'une grande partie du problème de suivi se fonde sur la reconnaissance de l'objet d'intérêt à partir d'un ensemble de caractéristiques visuelles.

1.2.3 Bilan des méthodes de détection et choix technique

Dans cette partie d'état de l'art, nous avons présenté les méthodes permettant la détection des OR. Nous allons maintenant faire le bilan afin de pouvoir synthétiser les différentes méthodes et orienter nos choix.
Les problèmes récurrents dans la détection d'obstacles routiers résident, d'un côté,

dans leurs grandes variabilités, de l'autre, dans les perturbations liées à l'utilisation des caméras embarquées en milieu urbain (changement de point de vues, les occultations partielles, ...).
Les variabilités des formes et des échelles des obstacles routiers posent certaines contraintes pour définir des régions d'intérêt. Un balayage de l'image à l'aide de fenêtres, ne serait pas une bonne solution dans la mesure où plusieurs tailles doivent être définies au préalable. En revanche, les ROI peuvent être définies en tirant profit des caractéristiques spécifiques des images IR où apparaissent des zones significatives de la présence de piéton. Une ROI peut contenir plusieurs obstacles routiers. Ce problème est surtout rencontré dans les situations de foule et d'occultations. Ainsi, il faudrait chercher dans les zones particulières de l'image des indices caractéristiques des obstacles à détecter. Les caractéristiques classiques issues des notions de symétries, de rapport hauteur/largeur ne sont pas assez robustes pour gérer le problème de variabilité et d'occultation. Pour résoudre ces problèmes, il faut intervenir dans le choix des caractéristiques présentant un fort pouvoir de généralisation. Cela ne peut être résolu qu'en utilisant des techniques d'apprentissage. Dans cette phase, un classifieur est utilisé afin de déterminer le modèle qui capture les attributs caractéristiques de chaque catégorie d'OR. Finalement, il est important de mentionner que les mêmes caractéristiques pourront être utilisées dans les différents processus de :

- Génération de ROI : pour la recherche d'indices et de primitives d'obstacles à détecter dans les zones particulières de l'image.
- Validation des hypothèses de détection : pour la classification.
- Le suivi temporel : pour la caractérisation d'un modèle d'objets à suivre.

Par conséquent, le problème de détection peut être traité comme étant un problème de représentation et de classification et ainsi de reconnaissance de catégories d'objet. Dans la section suivante, nous abordons cette problématique.

1.3 Problématique de la reconnaissance de catégories d'obstacles routiers

La reconnaissance de catégories d'objets est un processus cognitif important dans la perception, la compréhension des scènes et des objets et aussi dans la prise de décision. Chez l'être humain, la reconnaissance visuelle des scènes et des objets est généralement rapide, automatique et fiable. Autant dire que l'homme peut reconnaître des milliers de catégories d'objets instantanément et sans effort.

Cette preuve d'intelligence demeure spécifique à l'être humain, mais constitue un problème largement ouvert dans le domaine de vision par ordinateur. Ce problème réside dans la détermination de l'existence d'une catégorie spécifique d'un objet dans une image. Pour atteindre un tel objectif, il est indispensable de prendre en compte les apparences et les formes que peuvent avoir des objets au sein d'une catégorie. Ici, on considère trois catégories d'objets présents dans des scènes routières : véhicule, piéton et fond d'image. La première catégorie regroupe des objets rigides qui gardent toujours une forme rectangulaire. Lorsqu'un objet de type véhicule se déplace, tous les points qui le composent se déplacent de façon identique. Au contraire, les piétons sont des objets déformables présentant de grands changements de formes. Quant à la dernière catégorie, elle représente tout les objets, autres que véhicules et piétons qui peuvent être présents dans des scènes routières. La figure 1.5 expose quelques instances d'objets issus de ces trois catégories dans des images en VIS et en IR.

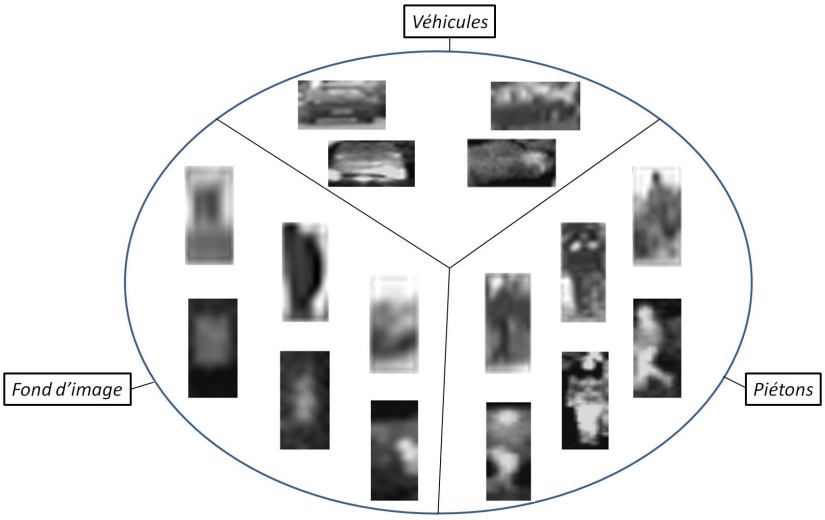


Figure 1.5. Quelques instances d'objets présents dans des images en Vis et en IR

Il est certain que les objets, appartenant à la même catégorie, ont des caractéristiques communes. Mais le principal problème qui se pose est comment identifier une signature compacte et pertinente pour chaque catégorie. Concrètement, le processus de reconnaissance de catégories d'objets passe par la mise au point d'un modèle de représentation pour pouvoir ensuite comparer les apparences ou les formes des objets. Le plus souvent, cette tâche de comparaison est confiée à un classifieur entrainé sur un ensemble d'apprentissage. Celui-ci opère sur un ensemble de caractéristiques extraites à partir d'un modèle de représentation. Ainsi,

la problématique se dessine clairement autour de deux problèmes majeurs : la représentation et la classification. Nous présenterons aussi le processus de fusion d'information qui permet d'enrichir la représentation et d'améliorer la prise de décision relative à la classification.

1.3.1 Problème de la représentation

En traitement d'image, la représentation d'objets dans les images est une tâche délicate. La difficulté réside dans la représentation du contenu de manière compacte et fidèle. En voulant reproduire un algorithme de détection, de reconnaissance ou de suivi, le problème suivant se produit en premier lieu : Comment représenter et reconnaître un objet ?
Pour ce qui est de la représentation structurelle de l'objet, la représentation consiste à définir une structure qui modélise l'objet. Par exemple, fenêtre englobante, maillage, graphe, ensemble de régions, etc. Pour ce qui est de la reconnaissance, cela revient à caractériser l'objet en calculant un ensemble de caractéristiques pour bien le distinguer des autres objets. La caractérisation des objets consiste, d'une façon générale, à extraire des descripteurs de forme, de texture ou bien d'apparence dans le but de représenter au mieux l'image en fonction de la tâche à réaliser.
Il est important, pour obtenir un modèle robuste de représentation, de couvrir autant que possible les transformations géométriques et photométriques que peut avoir un objet. Une bonne représentation émane d'une étude laborieuse de ces problèmes. Nous désignons par les transformations géométriques les transformations de type affine, les distorsions et les occultations partielles. Quant aux transformations photométriques, un bon modèle de représentation doit être robuste face aux changements de texture, couleur et d'illumination. Cette complexité s'intensifie quand il s'agit de représenter d'une façon robuste des OR à cause du mouvement non stationnaire de la caméra et des brusques changements d'illumination.

1.3.2 Problème de la classification

Le processus de classification consiste à reconnaître un objet en appariant ses caractéristiques avec celles des catégories connues des objets, afin de prédire la classe à laquelle l'objet pourrait se rattacher. Cette définition est donnée plutôt dans un contexte de classification supervisée qui consiste à induire, à partir

d'exemples étiquetés, une fonction associant une classe à un objet. Ce type de classification peut aussi être appelé analyse discriminante vu qu'elle permet de prédire l'appartenance à des catégories (ou classes) prédéfinies.
Les catégories considérées dans le cadre de notre étude ont déjà été définies (Piéton, Véhicule, Fond.). Il reste à définir la méthode de classification afin d'apparier ou de comparer les caractéristiques extraites d'un objet avec celles extraites à partir d'exemples étiquetés.
Le principe de la classification supervisée se base sur l'estimation d'une fonction de décision prédisant la classe d'une observation (une observation=vecteur de caractéristiques). Lorsque les distributions de probabilité des catégories ne sont pas connues dans l'espace de représentation, l'inférence des règles de décision est appelée apprentissage automatique de fonctions de décision. Lors de ce processus, l'algorithme apprend un modèle de classification à partir d'images labellisées, i.e. dont on connaît la classe. Ce modèle permet de prédire la classe d'une image non labellisée lors d'une phase de prédiction, appelée phase de test.
L'efficacité d'un tel algorithme est mesurée par sa capacité de généralisation sur un ensemble d'objets inconnus (jamais vus). Un manque de généralisation peut se traduire par le fait que l'algorithme ne commet pas d'erreurs sur les données déjà vues mais en commet beaucoup sur les autres. Ce phénomène est appelé surapprentissage. Empiriquement, cette situation pourrait être évitée en cherchant le minimum de l'erreur de généralisation sur un ensemble de validation. Bien évidemment, tout cela requiert que la distribution des observations dans l'espace de représentation contient suffisamment d'informations pour pouvoir à la fois estimer la fonction de décision et la valider sur un ensemble de validation. Une autre difficulté à souligner se rapporte aux données de grandes dimensions. Cette problématique influe considérablement sur la qualité et l'efficacité d'une technique de classification.

1.3.3 Fusion d'informations pour l'aide à la décision

La fusion de données consiste, d'une façon générale, à combiner différentes informations relatives à un problème. Une définition plus précise a été introduite par [Blo03] "La fusion d'informations consiste à combiner des informations issues de plusieurs sources afin d'améliorer la prise de décision". Dans le cadre de la reconnaissance des OR, la prise de décision consiste à combiner plusieurs sources d'informations afin d'identifier, d'une façon fiable, le type de l'obstacle routier. Ainsi, l'intérêt de la fusion est directement lié à l'amélioration des capacités de

décision.
Les sources d'informations peuvent provenir du même capteur, ou être issues de capteurs différents. L'application de la fusion de données issues du même capteur permet de tirer profit à la fois de la redondance et de la complémentarité des données. Ainsi, l'enjeu consiste à pouvoir extraire des sources d'information assez fiables et surtout complémentaires, en utilisant un seul capteur.
Dans la section 1.1.2.2, nous avons évoqué les problèmes posés par la fusion de capteur à savoir le temps de calcul, le coût et le calibrage. Bien que notre choix se soit porté a priori sur l'utilisation d'un système mono (voir section 1.1.2.3), nous n'écartons pas définitivement cette possibilité. Si les résultats obtenus en utilisant un seul capteur ne sont pas satisfaisants, il est intéressant d'étudier l'apport de la fusion de modalités VIS et INF.

1.4 Choix méthodologique

Après avoir positionné les principaux problèmes à résoudre, nous présentons dans cette section notre choix méthodologique et les bases des expérimentations.

1.4.1 Démarche adoptée

L'objectif majeur de cette thèse est d'analyser les processus conduisant à la conception d'un système de détection et de reconnaissance d'obstacles routiers, embarqué sur un véhicule. L'implémentation d'un tel système comporte de nombreux aspects, allant de la définition des régions d'intérêt aux techniques de reconnaissance et de détection.
Pour des raisons pratiques évidentes, nous avons scindé le système en une succession de processus. Ainsi découpé, le processus de détection et de reconnaissance comporte les aspects suivants :

1. la définition des régions d'intérêt et la génération d'hypothèses,
2. la représentation et la caractérisation des obstacles,
3. la classification et la reconnaissance du type de l'obstacle,
4. la fusion d'informations pour la prise de décision,
5. la validation de la détection de l'obstacle routier.

Tous de ces aspects sont dépendants les uns des autres. La définition des régions d'intérêt est un problème qui dépend fortement de l'aspect de la représentation et de la caractérisation de primitives images. Afin de dégager les caractéristiques

pertinentes à la discrimination des obstacles routiers, il convient d'évaluer les performances de la classification sur une base annotée d'apprentissage. La classification vise à étiqueter chaque objet en l'associant à une classe en se basant sur le modèle de représentation et sur l'ensemble des caractéristiques extraites. Quant au processus de fusion, il interviendra pour assurer l'amélioration de l'analyse et de l'interprétation des données fournies.
Dans ce manuscrit nous avons choisi de développer ces différents aspects, non pas en suivant les étapes ordonnées du processus de détection, mais plutôt en respectant leurs dépendances. La relation entre tous les aspects liés à la problématique de thèse sont décrits par la figure 1.7. Cette figure donne un aperçu de la structure de la thèse. Cette structure déploie la problématique et les aspects liés aux différents chapitres. Dans le chapitre suivant, nous faisons le point sur l'application des techniques de représentation, de classification et de fusion des données au problème de reconnaissance des obstacles routiers. Un état de l'art est dressé et les différentes méthodes proposées sont comparées et analysées. C'est à partir du troisième chapitre que l'on commence à présenter nos contributions. Un modèle de représentation locale et globale est proposé permettant de caractériser à la fois l'apparence locale, la forme et la texture des objets routiers. L'apport de la fusion de données capteurs est évalué dans le quatrième chapitre. Finalement, les conclusions tirées des résultats obtenus en classification et en fusion seront mis à profit pour la conception d'un système de détection de piétons. Le dernier chapitre en fera une étude détaillée et exposera les résultats obtenus de l'expérimentation sur des données de piétons. Les bases d'images utilisées sont présentées dans la section suivante.

1.4.2 Description des bases d'images utilisées

Toutes les études présentées dans cette thèse sont validées par des expérimentations importantes sur des bases d'images en IR lointain et en VIS. Ces images nous ont été fournies par le laboratoire italien VisLab [4]. Les images utilisées ont été extraites par un système appelé Tetravision [BBFV06]. Ce système de vision comprend deux paires stéréo IR et VIS. Ainsi, dans cette base on trouve des images issues de caméra VIS et leurs homologues de caméras IR. Néanmoins, nous avons mené nos expérimentations seulement sur des images en mono ayant la résolution de 320×240 pixels. Les caméras infrarouges utilisées sont sensibles aux longueurs des ondes se situant entre 7 et 14μm. Des images représentatives des bases d'images

4. http ://en.wikipedia.org/wiki/VisLab

utilisées sont illustrées dans les figures 1.6.

Figure 1.6. Quelques exemples d'images VIS et IR de la base de Tetravision

Pour réaliser nos expérimentations, nous avons annoté manuellement 3917 objets provenant de quatre bases d'images nommées :

- *Tetra1* : contient 366 piétons.
- *Tetra2* : contient 455 piétons.
- *Tetra5* : contient 2111 piétons.
- *Tetra6* : contient 986 obstacles routiers.

Les bases Tetra1,Tetra2,Tetra5 ont été utilisées afin de réaliser des expérimentations sur la détection et la reconnaissance des piétons. Les deux premières bases (Tetra1 et Tetra2) ont servi en tant que bases de test. La base Tetra5 a été divisée en deux parties égales, l'une servant à l'apprentissage (base d'apprentissage) et l'autre à la validation (base de validation). Tandis que, la base Tetra6 a été utilisée afin d'expérimenter le système de reconnaissance multiclasses. Notons que les différentes images utilisées sont prises en conditions réelles, de jour et dans un milieu urbain.

1.5 Conclusion

Dans ce chapitre, nous avons analysé les difficultés liées aux applications de détection et de reconnaissance des obstacles routiers d'un système de monovision embarqué. Cette analyse nous a permis non seulement d'identifier les problématiques, mais aussi de mettre en évidence un ensemble d'aspects différents : la représentation, la classification et la fusion d'information. L'état de l'art de ces différents aspects est présenté dans le chapitre suivant.

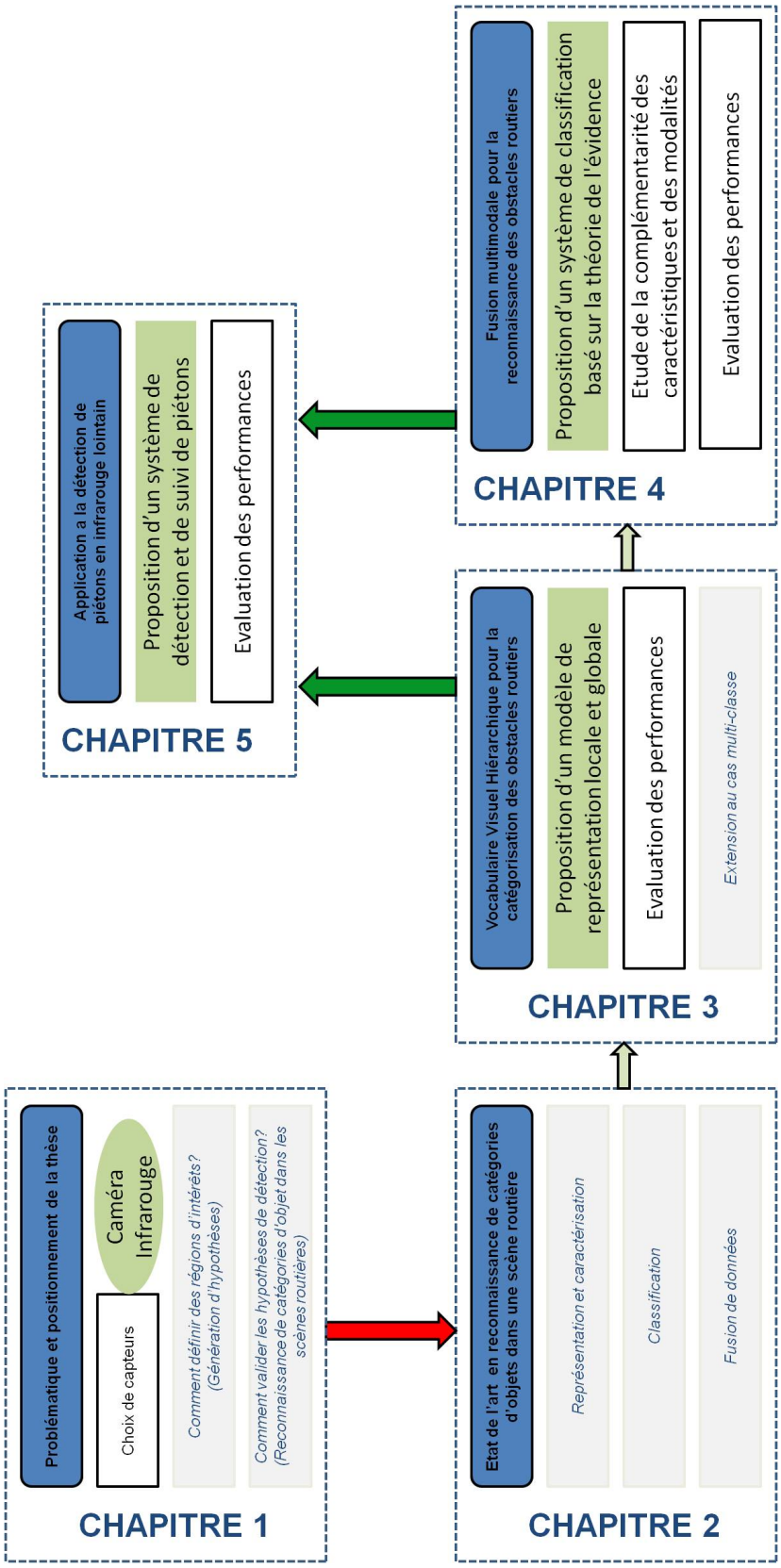


Figure 1.7. Problématique et structuration de la thèse

Chapitre 2

Reconnaissance de catégories d'objets dans les scènes routières

Introduction

Dans le chapitre précédent, nous avons mis en évidence le concept de reconnaissance de catégories d'objets dans une scène routière pour parvenir à implémenter un bon module de détection des obstacles routiers. Nous avons également brièvement présenté les techniques de base de la reconnaissance, à savoir la représentation, la classification et la fusion d'informations.
Dans ce chapitre, nous faisons le point sur l'application de ces techniques au problème de reconnaissance des obstacles routiers. Un état de l'art est dressé et les différentes méthodes proposées sont comparées et analysées.

2.1 Représentation des obstacles routiers

Il existe de nombreuses manières de constituer un état de l'art sur les méthodes de représentation et de caractérisation des objets. Toutefois, nous cherchons à faire un tour d'horizon des principales méthodes proposées dans le cadre de ce travail de thèse, à savoir la détection et la reconnaissance des OR.
Très peu d'auteurs [Sua06] ont interprété le problème comme étant un problème de représentation structurelle. La majorité ont cerné la position générale du problème en recensant l'ensemble des caractéristiques visuelles qu'il faut extraire afin de représenter une telle catégorie d'objets. Dans ce paragraphe, nous regroupons en trois approches les méthodes proposées afin de caractériser visuellement les OR :

globale, par région et locale.

2.1.1 Approche globale

La caractérisation globale consiste à décrire l'intégralité d'une image par des caractéristiques calculées en utilisant tous les pixels de cette image. L'exemple le plus connu à cet égard est celui des histogrammes qui représentent la distribution des intensités de l'image. En littérature, il existe de très nombreux types de caractéristiques que nous essayons de grouper en deux familles. Dans la première, le critère global résulte de la combinaison des caractéristiques qui peuvent être extraites localement (des points de contours, des unités de texture). Tandis que dans la deuxième famille, chaque caractéristique décrit, par elle même, l'image globalement.

2.1.1.1 Représentation globale par des caractéristiques extraites localement

Nous présentons dans cette section les descripteurs les plus souvent utilisés à savoir, l'image des contours, les histogrammes des différences d'intensité et le numéro de l'unité de texture.

L'image des contours

L'extraction des contours d'une image permet de repérer les points qui correspondent à un changement brutal de l'intensité lumineuse. L'application d'un filtre optimal, comme celui de Canny [Can86], permet d'une part, de réduire de manière significative la quantité des données et d'autre part, d'éliminer les informations qui sont moins pertinentes. Le résultat de ce processus est l'obtention d'une image binaire, c'est à dire constituée de 0 et de 1. La concaténation de l'ensemble des valeurs de pixels est ensuite directement appliquée à l'entrée d'un classificateur. De même, il est possible d'envisager une représentation plus précise en attribuant à chaque point de contour une valeur non binaire comme la valeur du gradient ou de son orientation. Toutes ces méthodes permettent de générer un vecteur de caractéristiques ayant la même taille que l'image d'intérêt.

Histogramme des différences d'intensité

La détermination de l'histogramme est réalisée en calculant les différences d'intensités entre les pixels voisins selon une orientation bien définie. Cela permet de

générer un vecteur de caractéristiques dont la taille est donnée par le produit entre le nombre d'orientations et le nombre des niveaux de gris.

Numéro de l'unité de texture

Cette méthode propose de décomposer une image en un ensemble d'unités appelées unités de textures (NTU). Considérons un voisinage de 3*3, les valeurs des huit éléments entourant le pixel central sont remplacées par les valeurs : 0, 1 ou 2. À cette unité de texture est alors associé un label calculé en fonction de ces valeurs. Cette méthode permet de générer un vecteur de caractéristiques ayant la même taille que l'image d'intérêt.

2.1.1.2 Les descripteurs globaux

Les descripteurs globaux les plus utilisés sont des descripteurs statistiques qui caractérisent la texture des images. Généralement, ces descripteurs sont déterminés à partir des matrices de cooccurrence, ou suite à un filtrage fréquentiel, ou à partir des statistiques du premier ou d'ordre élevé. L'ordre des statistiques se calcule en fonction du nombre de pixels mis simultanément en jeu. Par exemple, la moyenne ou la variance des intensités dans une image sont du premier ordre. En revanche, les statistiques extraites en comptant le nombre de transitions entre deux intensités lumineuses, sont d'ordre deux. Ces statistiques peuvent être calculées en analysant l'image directement ou après une étape de filtrage. Nous parlons ici de filtrage préalable qui consiste à explorer le domaine fréquentiel en analysant les fréquences spatiales (Fourrier, DCT, ondelettes, etc) afin de retrouver une trace du motif de la texture. L'exemple le plus caractéristique à cet égard est celui du filtre de Gabor qui a été largement utilisé dans la littérature afin de caractériser les véhicules [SBM02, SBM05, SBM06] et les piétons OR [CZQ05]. Dans [SBM02], l'auteur a montré que ce filtre est efficace vu qu'il présente une invariance à l'intensité, une sélectivité à l'échelle et à l'orientation. La réponse impulsionnelle d'un filtre de Gabor ($g(x, y)$) se définit comme une fonction gaussienne modulée par une sinusoïde :

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{{\sigma_x}^2}+\frac{y^2}{{\sigma_y}^2}\right)} e^{-i2\Pi(u_0 x + v_0 y)}, \tag{2.1}$$

où σ_x et σ_y sont respectivement les écarts types de la modulation Gaussienne dans les directions spatiales x et y.
La plupart du temps, les valeurs calculées dans les matrices de cooccurrence ou

obtenues après un filtrage fréquentiel ne sont pas utilisées directement comme descripteurs. En effet, on en déduit des caractéristiques plus compactes représentant généralement des moments statistiques. Dans le tableau 2.1, nous présentons les caractéristiques les plus courantes à cet effet. Nous considérons dans ce tableau que ces caractéristiques se déduisent de la probabilité empirique $p(n)$ du niveau de gris n.

Table 2.1. Quelques moments statistiques

Caractéristiques	**Calcul**
Les moments d'ordre k	$\mu_k = \sum^{n} n^k p(n)$
La moyenne	μ_1
La variance	$\sigma^2 = \mu_2$
Le biais	$\frac{\mu_3}{\sigma^3}$
L'applatissement (kurtosis)	$\frac{\mu_4}{\sigma^4} - 3$
L'énergie	$W = \sum^{n} p^2(n)$
L'entropie	$E = -\sum_{n} p(n) \log p(n)$
Le contraste	$\frac{max(n)-min(n)}{max(n)+min(n)}$

L'approche globale est connue par sa rapidité et sa simplicité de mise en œuvre. En suivant cette approche, des travaux récents ont été menées [ARB09b, ARB09a, Dis10] afin de caractériser les OR. Ces travaux ont montré que la combinaison de plusieurs caractéristiques globales peut parvenir à obtenir de bons résultats.
Cependant, l'approche globale souffre de plusieurs problèmes. En effet, elle suppose implicitement que la totalité de l'image soit reliée à l'objet. Ainsi, tout objet incohérent introduirait du bruit dans les caractéristiques. Cette limitation incite de fait à se tourner vers des méthodes par région, voire locales.

2.1.2 Approche par région

L'approche par région consiste à décomposer l'image en nombreuses régions de tailles fixes ou variables avant de caractériser chacune d'entre elles. Cette décomposition se fait d'une manière généralement prévisible afin que les caractéristiques extraites soient homogènes entre elles.
Le découpage en régions peut être fait de deux façons. La première consiste à définir préalablement des régions adjacentes de tailles fixes et d'y extraire le même type de caractéristiques. Cette méthode est plus générique car elle peut être employée pour détecter n'importe quelle catégorie d'objet. La deuxième méthode

de découpage peut être appelée multiparties car elle définit des régions d'intérêt spécifiques à chaque partie de l'objet.

2.1.2.1 Caractérisation visuelle de régions fixes

Dans le cadre de la caractérisation des OR par des régions de tailles fixes, les deux références à ce sujet sont les pseudo-Haar (Haar-like features en anglais) [OPS+97, HK09] et les histogrammes d'orientation de gradients [DT05, BBDR+07, CSY08]. Ces deux caractéristiques ont montré de bonnes performances pour la caractérisation des véhicules ainsi que des piétons.

Les caractéristiques pseudo-Haar

La caractérisation par les pseudo-ondelettes de Haar [OPS+97, PTP98] reste parmi les méthodes les plus connues en littérature pour la détection d'objets dans les images. Elle a été utilisée dans le premier détecteur de visages en temps réel [VJ02] et fut la première méthode proposant d'appliquer un classifieur SVM [Vap95] pour la détection de piétons. L'avantage principal de ces caractéristiques est la rapidité de leur calcul. Elles permettent de représenter la forme générale contenue dans l'image très rapidement en utilisant une technique d'image intégrale.
Plusieurs configurations d'ondelettes ont été proposées [VJ02, LM02]. Les configurations les plus utilisées permettent de caractériser les contours dans les trois directions possibles : horizontalement, verticalement et diagonalement. Ces configurations sont illustrées dans la figure 2.1.

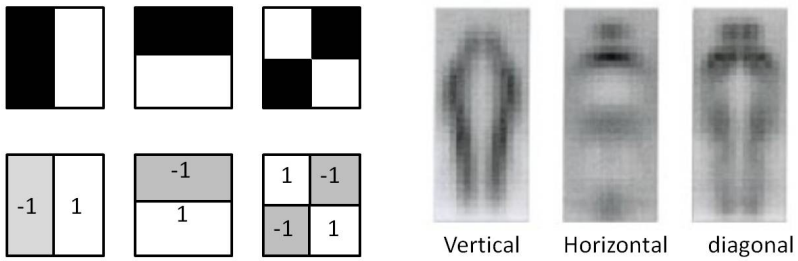


Figure 2.1. Les trois configurations principales d'ondelettes de Haar et leurs résultats de filtrage pour une image de piéton.

L'image est donc caractérisée par trois ensembles de coefficients issus de la décomposition en ondelettes. Chaque ensemble de coefficients est calculé en utilisant des fenêtres qui délimitent des zones rectangulaires adjacentes. Ensuite, les intensités de pixels des deux zones (sombre et claire, dépendent du type du filtre

utilisé) sont additionnées, formant deux sommes dont la différence constitue une caractéristique. D'où l'obtention d'une matrice de coefficients, dont les éléments seront ensuite concaténés afin de former un vecteur de caractéristiques.
Cette méthode a démontré son efficacité en terme de caractérisation pertinente de l'image et reste ainsi une référence dans le domaine de la détection des OR.

Les Histogrammes de gradient orienté

Les histogrammes de gradient orienté (HOG) sont des histogrammes locaux de l'orientation du gradient calculés sur des régions régulièrement réparties sur l'image [DT05]. La concaténation de ces histogrammes permet de définir un vecteur caractéristique.
En littérature, les HOG ont été utilisés aussi bien que les caractéristiques pseudo-Haar. Dans le cadre de la détection des OR, les travaux faisant l'objet de comparaison ont montré que ces caractéristiques sont plus discriminantes que les caractéristiques de pseudo-Haar [LCL09, SAM+09].
Les vecteurs caractéristiques sont calculés systématiquement par le même procédé. Tout d'abord, un angle de gradient est calculé pour chaque pixel. Ensuite, l'image est découpée en des régions de taille fixe (Comme le montre la figure 2.2). Généralement, la taille des cellules est fixée au préalable selon les besoins et les performances obtenues. Pour chaque cellule, un histogramme d'orientation de gradient est calculé par l'accumulation des votes des pixels. La dernière étape consiste à normaliser chaque histogramme afin de contourner les problèmes de changements d'illumination.

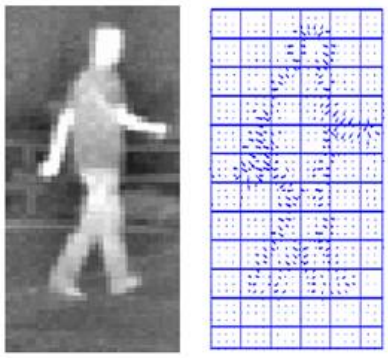


Figure 2.2. Exemple de découpage d'une image avant d'être caractérisée par les HOG

Les HOG sont des caractéristiques très pertinentes. Elles sont les plus couramment utilisées pour caractériser et détecter les OR. Néanmoins, le découpage de

l'image est fait selon un modèle rigide qui suppose la présence d'un piéton sans posture particulière (debout, non occulté, bien centrée dans l'imagette). De plus, le problème d'occultation n'est également pas complètement résolu, sauf en envisageant dans l'ensemble d'apprentissage plusieurs images contenant des piétons masqués partiellement (occultés).

2.1.2.2 Extraction des caractéristiques multiparties

Cette méthode se distingue de la première approche par régions fixes puisqu'elle définit des régions d'intérêt spécifiques à chaque partie de l'objet. Cette méthode est utilisée surtout pour la caractérisation des piétons dans les images [SGH04, ASDT$^+$07, GM07]. Nous illustrons à titre d'exemple dans la figure 2.3 deux méthodes de découpage qui respectent la spécificité des parties du corps d'un piéton.

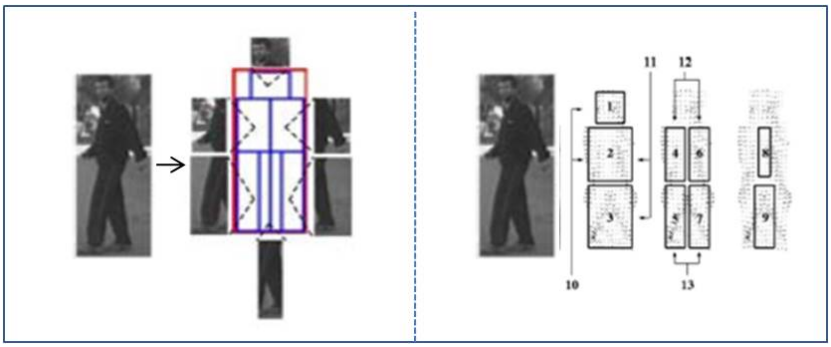


Figure 2.3. Deux exemples de découpage d'images en régions proposés, respectivement, dans [ASDT$^+$07] et [SGH04]

La principale motivation derrière cette méthode est de considérer qu'un objet est constitué de différentes parties non homogènes devant ainsi être caractérisées et détectées par des algorithmes différents. Les principales caractéristiques citées jusqu'à maintenant : DCT, Gabor, matrice de cooccurence, Haar, HOG, l'image de contours, NTU, ... peuvent être utilisées pour caractériser des parties spécifiques de l'objet. Dans [ASDT$^+$07], l'auteur a adopté cette méthode en testant plusieurs types de caractéristiques pour chaque partie. Pour une représentation du piéton, il montre que les caractéristiques NTU sont très performantes pour la tête, les bras et la zone située entre les jambes. Tandis que, l'image des contours convient plus pour caractériser les jambes.
La caractérisation multiparties est très intéressante au point de vue discrimination, notamment pour les objets déformables. Plusieurs travaux ont montré qu'un

détecteur d'objet déformable qui procède par parties est plus efficace. Mais la problématique posée ici, concerne le choix des caractéristiques à employer pour représenter une telle partie de l'objet.
Une solution alternative très prometteuse consiste à extraire les mêmes caractéristiques pour plusieurs parties et à utiliser une méthode de boosting afin d'en sélectionner les meilleures. Nous reviendrons sur la technique de boosting dans la section 2.2.1. Viola et Jones ont adopté cette solution [VJ02] et leur méthode fait partie des toutes premières méthodes capables de détecter efficacement et en temps réel des objets dans une image [1]. Depuis, plusieurs travaux reprenant le même principe ont été menés mais en employant des caractéristiques autre que les pseudo-Haar. Parmi ces travaux, nous citons plus particulièrement ceux qui ont montré de bonnes performances dans le cadre de la catégorisation des OR [WN06, WN07, TPM08, PSZ08]. Dans ce qui suit, nous détaillons ces méthodes.

Les Edgeletes

Les edgeletes sont des caractéristiques locales extraites depuis la silhouette de l'objet. Elles représentent de courts segments de lignes ou de courbes dans l'image de contours de l'objet. Les Edgeletes ont été utilisées pour la première fois par Wu et Nevatia [WN05] dans le cadre de la détection de piétons. Les bons résultats obtenus ont été confirmés dans des travaux ultérieurs [WN06, WN07].

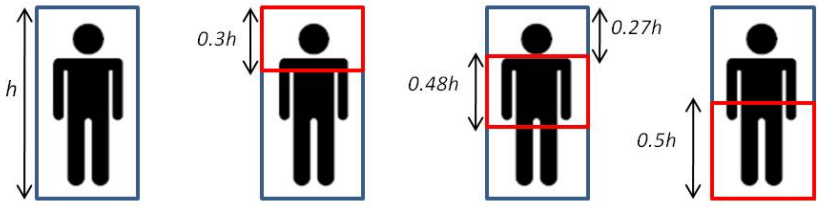


Figure 2.4. Découpage des parties du corps du piéton en tête-épaules, torse et jambes

Comme le montre la figure 2.4, le découpage en régions se fait selon trois parties du corps : tête-épaules, torse et jambes. Ce type de modélisation permet de détecter les parties du corps indépendamment les unes des autres. Le fonctionnement par parties assure évidemment une robustesse aux occultations. Néanmoins, la constitution d'une base d'apprentissage multiparties demande beaucoup d'ef-

1. http ://fr.wikipedia.org/wiki/Méthode_de_Viola_et_Jones

forts dans la mesure où chaque partie de l'objet doit être annotée manuellement. Dans la suite, nous présentons la méthode proposée dans [TPM08] qui permet de dépasser cette limite en sélectionnant, dans le processus d'apprentissage, les meilleures régions à considérer.

La covariance de région

La covariance de région rassemble un ensemble de 8 caractéristiques issues d'une matrice de covariance d'emplacement, d'intensités lumineuses et de gradients. Ces caractéristiques sont calculées localement sur une région d'intérêt à l'aide d'une image intégrale.
Les caractéristiques de covariance ont d'abord été introduites dans [TPM06] pour faire l'appariement et la classification de texture. Ensuite elles ont été appliquées à la détection de piétons [TPM08]. Dans cette dernière référence, le piéton est représenté par plusieurs régions, de tailles variables, et qui peuvent se chevaucher, la figure 2.5 en illustre le principe. Chaque région est caractérisée par des descripteurs de covariance. Les régions les plus pertinentes sont sélectionnées par une méthode de recherche gloutonne. Enfin, la classification est réalisée grâce à plusieurs cascades de classifieurs boostés. Nous revenons dans la section 2.2.1 sur le modèle de cascade de classifieurs.

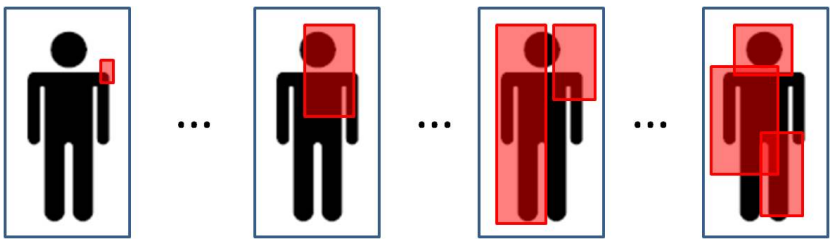


Figure 2.5. Génération de plusieurs descripteurs de covariance pour chaque image. L'image est parcourue dans tous les sens et avec des fenêtres de tailles différentes. Les régions les plus pertinentes sont ensuite sélectionnées par une méthode de recherche gloutonne.

L'approche par région calcule une signature intégrant l'influence de tous les pixels. Ainsi, elle ne permet pas de surmonter les problèmes d'occultations prononcées ou ceux liés à la présence d'un fond chargé. Dans ces situations difficiles, il convient de ne considérer que des zones qui sont rarement occultées et qui ne subissent pas d'influence de fond chargé. C'est là tout l'intérêt de l'approche locale.

2.1.3 Approche locale

Les techniques de représentation d'objets par une approche globale ou par régions ont reçu une attention décroissante dans la communauté scientifique depuis l'avènement des approches locales et notamment des points d'intérêt (On notera désormais POI les points d'intérêt). En effet, la représentation d'objet par un ensemble de POI est devenue un outil performant pour répondre aux problèmes posés par les larges variations de formes et d'apparences, ainsi qu'aux problèmes d'occultations partielles. Ces dernières années, plusieurs travaux de recherches ont été entrepris dans ce type de représentation, décrivant différentes méthodes de détection et de caractérisation locale [HS88, Low99, SM97, Low04, BTG06] par POI.
Le principe de la caractérisation locale d'une image se fonde sur l'identification des POI, puis sur l'utilisation de descripteurs locaux qui, par opposition aux descripteurs globaux, ne caractérisent qu'une zone restreinte de l'image. Dans ce qui suit, nous citons les principales méthodes proposées pour détecter et caractériser les POI.

2.1.3.1 Détection de POI

En littérature, il existe différentes définitions du terme point d'intérêt. D'autres termes spécifiques sont souvent employés : coin, point stratégique, point particulier, etc. De manière générale, il s'agit de points riches en termes d'informations sélectionnées selon un critère précis, et qui peuvent mieux décrire un objet que d'autres.
L'objectif de la détection de POI est de localiser les points les plus importants de l'objet présentant une forte variabilité dans le signal visuel. De nombreuses approches ont été expérimentées, mais nous nous limitons à citer les plus populaires. On distingue deux types de détecteurs : *à échelle fixe* et *multiéchelles*. Le terme multiéchelles est employé quand la réponse du détecteur est calculée avec des tailles de fenêtre ou des résolutions d'images différentes.
Les détecteurs de *Beaudet* et *Moravec* [Mor77, Bea78] furent les premiers algorithmes de détection de POI. L'opérateur de Beaudet [Bea78] s'appuie sur l'extraction des maximas locaux après le calcul de la dérivée seconde pour chaque pixel de l'image. L'opérateur proposé car Moravec [Mor77] s'appuie, quant à lui, sur une matrice d'auto-corrélation d'une petite zone de l'image. Celle-ci va quantifier les différences de niveau entre la zone considérée et la même zone translatée

dans quatre directions (pour plus de détails sur la méthode, voir [Kra08]). Si le minimum d'une de ces quatre valeurs est supérieur à un seuil, alors il s'agit d'un point d'intérêt.
Les détecteurs de *Beaudet* et *Moravec* [Mor77, Bea78] furent les premiers algorithmes de détection de POI. L'opérateur de Beaudet [Bea78] s'appuie sur l'extraction des maxima locaux après le calcul de la dérivée seconde pour chaque pixel de l'image. L'opérateur proposé car Moravec [Mor77] s'appuie, quant à lui, sur une matrice d'auto-corrélation d'une petite zone de l'image. Celle ci va quantifier les différences de niveau entre la zone considérée et la même zone translatée dans quatre directions (pour plus de détails sur la méthode, voir [Kra08]). Si le minimum d'une de ces quatre valeurs est supérieur à un seuil, alors il s'agit d'un point d'intérêt.
Harris et Stephen [HS88] ont identifié certaines limitations liées à la réponse de ce détecteur. Ils en ont déduit un détecteur de coins très populaire : le détecteur de *Harris*. Ce détecteur, simple à mettre en œuvre et rapide, a été le plus utilisé dans les années 1990. Cependant, cet opérateur n'a pas été adapté aux changements d'échelle. Afin de faire face à ces transformations, Mikolajczyk et Schmid [MS04] ont proposé une représentation multiéchelles pour ce détecteur. Ainsi, les POI détectés doivent également être des maxima dans l'espace-échelle du laplacien. Considérons que les niveaux successifs d'échelles sont représentés par σ_n, ces points sont sélectionnés en appliquant l'équation suivante :

$$|LOG(x, \sigma_n)| = {\sigma_n}^2 |L_{xx}(X, \sigma_n) + L_{yy}(X, \sigma_n)| \tag{2.2}$$

L'équation 2.4 mesure la réponse du filtre Laplacien de Gaussienne (LOG) pour un point $X = (x, y)$. Avec L_{xx}, désigne le résultat de la convolution de l'image avec la dérivée seconde par rapport à x. Ensuite l'échelle est évaluée en représentant l'image comme une pyramide telle que chaque niveau correspond à une échelle. Concrètement, la convolution par gaussienne revient à lisser l'image afin de lutter contre l'apparition de points fictifs. Les zéros du laplacien caractérisent les points d'inflexions de l'intensité et déterminent, dans une certaine mesure, la présence de coins ou de contours.
La représentation multiéchelles peut être déterminée en se reposant sur l'idée que le laplacien peut être vu comme la différence entre deux lissages gaussiens de tailles différentes. Cette méthode, appelée DOG (Différence of Gaussians), constitue une bonne alternative au LOG (Laplacian of Gaussian) pour accélérer le calcul d'une représentation.

$$|DOG(x, \sigma_n)| = |I(x) * g(\sigma_n) - I(x) * g(k\sigma_n)| \tag{2.3}$$

Avec $*$ désignant le produit de convolution, $g()$ une fonction gaussienne d'écart type σ_n et I l'image.
Les Dogs sont un des piliers de la méthode proposée par Lowe [Low04], appelée *SIFT*(Scale-Invariant Feature Transform) et qui se révèle la plus populaire. La phase de détection dans l'algorithme de SIFT se repose sur la construction d'une série d'images approximant l'espace d'échelle associé à une image. Les POI sont extraits en cherchant les extrema locaux autour de 26 voisins directs de chaque pixel (9 au dessus et respectivement en dessous et 8 au même niveau). Afin d'améliorer la localisation des POI dans l'escape et l'échelle, une étape d'interpolation est utilisée fournissant des coordonnées sub-pixelliques du point. Cette étape de détection est extrêmement sensible donnant un nombre trop important de points. Ainsi, les points de faible contraste sont purement éliminés. Dans le même but, une méthode itérative est employée pour faire converger les POI qui ne se situent pas au niveau de l'extremum local. Enfin, les valeurs propres 2×2 de la matrice hessienne [2] sont analysées afin de différencier les points de contour des coins. Bien qu'il soit couteux en temps de calcul, l'algorithme SIFT a acquis une importance considérable auprès de la communauté de vision par ordinateur. Etant donnée ces performances considérables, plusieurs études se sont penchées sur des améliorations possibles surtout au niveau du temps de calcul. Afin d'accélérer le processus de description des POI détectés, Ke and Sukthankar [KS04] ont proposé le PCA-SIFT qui emploie la technique d'analyse en composante principale afin d'accélérer les temps d'appariement des POI SIFT. Par l'utilisation d'une structure de données adaptée, Grabner et al. [GGB06] affirment gagner un facteur de 8 en vitesse, avec une légère baisse de performances. Mais à ce sujet, nous considérons que les alternatives proposées dans [BTG06] constituent les améliorations les plus efficaces à apporter surtout en temps de calcul. Dans [BTG06] Bay et al. ont proposé le détecteur *SURF* (Speedep Up Robust Features) qui s'appuie sur l'approximation de la matrice Hessienne et l'utilisation de l'image intégrale. Ce détecteur présente un bon compromis entre robustesse au changement d'échelle et temps de calcul.
La matrice hessienne $H(\sigma, x)$ pour une échelle σ est définie comme suit :

$$H(\sigma, X) = \begin{pmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{yx}(X, \sigma) & L_{yy}(X, \sigma) \end{pmatrix}$$

2. La matrice hessienne représente la matrice carrée contenant les dérivées partielles secondes

L'approximation de $L..(X, \sigma)$, représentant la convolution de l'image avec la dérivée seconde, se base sur l'application des masques de convolution. Ainsi, l'espace des échelles est estimé en appliquant des filtres de tailles différentes tout en tenant compte de la nature discrète de l'image. Le premier filtre appliqué est de taille $9 * 9$, ensuite les couches suivantes sont obtenues par le filtrage de l'image avec des masques plus grands ($15 * 15$, $21 * 21$ et respectivement $27 * 27$).
La réponse du détecteur est basée sur le déterminant de la matrice hessienne. En effet, les points détectés doivent également être des maxima dans l'espace-échelle du déterminant de la matrice hessienne. Concrètement, le déterminant de la matrice représente la puissance de la réponse de la région (blob) considérée autour du POI détecté. Pour une position (x, y), ce dernier est estimé en utilisant l'équation suivante :

$$det(H_{approx}) = D_{xx}D_{yy} + (0.6D_{xy})^2 \geq seuil \tag{2.4}$$

Avec D_{xx}, D_{yy} et D_{xy} désignant les résultats de l'application des filtres. Le seuil est une constante déterminée de façon empirique. Afin d'améliorer la localisation des POI dans l'espace et l'échelle, semblablement à l'algorithme SIFT, une étape d'interpolation est utilisée.
Toutes ces approximations permettent, sans doute, un gain considérable en temps de calcul. Plusieurs études comparatives des performances des algorithmes SURF et SIFT ont été réalisées [BP07, ETLF11]. Ces études ont montré que bien qu'une implémentation du SURF puisse améliorer nettement les temps de calcul, les performances de SIFT restent légèrement supérieures. Les critères de performances ont été établis en fonctions de la robustesse des descripteurs devant différentes transformations d'images. Dans la section suivante, nous présentons ces différentes transformations ainsi que les descripteurs spécifiques aux détecteurs SIFT et SURF.

2.1.3.2 Les descripteurs de POI

Un descripteur local d'un POI fournit une caractérisation locale sous la forme d'un vecteur d'attributs. Le caractère le plus marquant d'un descripteur est sa robustesse face aux transformations usuelles. Cette propriété révèle de la capacité à établir le même descripteur (ou très similaire) pour le même point quelque soit les transformations que peut subir l'image. Les principales transformations sont :

- Le changement d'illumination : il peut résulter de plusieurs facteurs qui sont liés directement au type de capteur ou le plus souvent aux conditions d'éclairage.

En effet, les capteurs ont des sensibilités différentes aux conditions de faible ou de forte illumination. En ce qui concerne les conditions d'éclairage, elles sont liées aux changements des conditions d'acquisition de l'image qui dépendent à leur tour, du temps, des angles de vues et de la nature de la scène routière.

- La rotation : elle est due essentiellement au bougé lors de la prise de vue.

- Le changement d'échelle : il se traduit par la modification de la résolution de l'image contenant un objet. Cette variation entraine généralement une perte de précision dont l'amélioration n'est pas évidente en adaptant la résolution spatiale (zoom).

- Le changement du point de vue : ce facteur peut conduire au changement de la composition de l'image. En effet, la forme d'un objet ne sera pas la même s'il est acquis selon deux points de vue différents. Il est à noter aussi que lors d'un changement de point de vue, la variation d'échelle n'est pas forcément uniforme mais variable d'une direction à une autre.

Les algorithmes SIFT et SURF présentent des propriétés d'invariance par rapport à ces transformations. En effet, ces algorithmes ne permettent pas seulement de détecter des points d'intérêt invariants à l'échelle mais bien aussi de construire des descripteurs robustes à plusieurs transformations.

Les descripteurs SIFT

Pour l'algorithme SIFT, les étapes de détection, détaillées précédemment, se résument en deux phrases : la détection des extrema des DOG pour une même échelle et la vérification de la stabilité sur plusieurs échelles. Ainsi chaque POI détecté se voit attribuer une valeur d'échelle spécifique. Cette valeur est utilisée par la suite pour fixer la taille de la région (voisinage) considérée pour caractériser le POI. Afin d'assurer une invariance à la rotation, le point détecté se voit attribuer une orientation principale, selon laquelle, un changement des coordonnées locales au voisinage du POI est effectué. La valeur de l'orientation principale correspond au maximum détecté dans un histogramme contenant les valeurs de l'orientation du gradient de tous les pixels voisins.
Ensuite 16 histogrammes locaux, représentant l'orientation locale du gradient sur des zones de $4*4$ pixels autour du point central, sont établis. Chaque histogramme contient 8 bins qui représentent les 8 orientations principales entre 0 et 360 degrés. Par la suite, les histogrammes obtenus sont normalisés afin d'assurer une invariance aux changements d'illumination. On obtient finalement des descripteurs SIFT ayant une dimension égale à $4 * 4 \times 8 = 128$ descripteurs. La figure 2.6

illustre le principe d'extraction de descripteurs SIFT autour d'un POI.

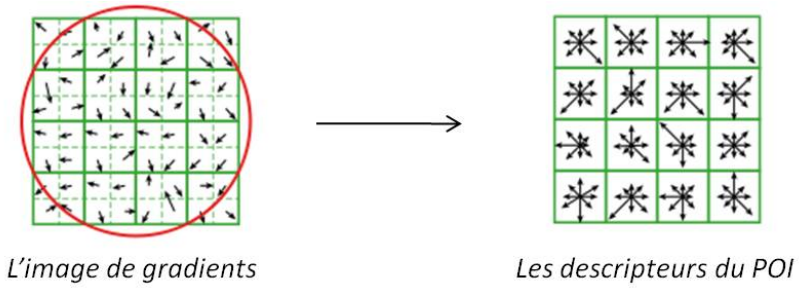


Figure 2.6. Les descripteurs SIFT d'un POI

Les descripteurs SURF

À la différence du SIFT qui utilise les HOG pour décrire les points d'intérêt, le SURF se base sur le calcul des sommes de réponses d'ondelettes de Haar. Les réponses sont représentées par des points dans l'espace. L'orientation locale est calculée en sommant les réponses verticales et horizontales inclues dans une zone de taille $\Pi/3$, comme le montre la figure suivante :

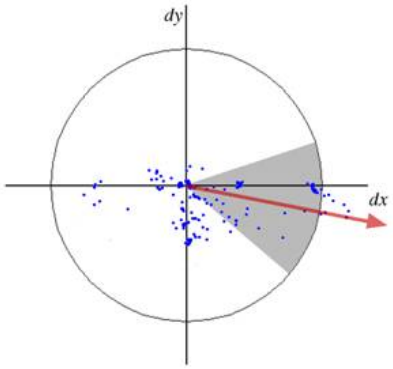


Figure 2.7. Détermination de l'orientation principale d'un POI SURF

Afin d'obtenir une invariance à la rotation et à l'échelle, l'algorithme reprend les mêmes techniques que SIFT. Après la détermination de la valeur d'échelle et de l'orientation principale du POI, une région d'intérêt est découpée en bloc de 4×4. Dans chaque bloc des descripteurs simples sont calculés formant un vecteur v défini par :

$$v = (\sum dx, \sum dy, \sum |dx|, \sum |dy|) \tag{2.5}$$

Avec dx (et respectivement dy) sont les réponses d'une analyse par ondelettes de

Haar dans la direction horizontale (et respectivement verticale). Cela conduit à l'obtention d'un vecteur de descripteurs ayant une dimension de $4 \times 4 \times 4 = 64$ (SURF-64). De même, il est possible de construire un descripteur de 128 éléments (SURF-128) en calculant les termes suivants de manière séparée :

- $\sum d_x$ et $\sum |d_x|$ pour $d_y < 0$ et $d_y \geq 0$
- $\sum d_y$ et $\sum |d_y|$ pour $d_x < 0$ et $d_x \geq 0$

Notons pour conclure que la recherche sur la détection et la description des POI est toujours très active et de nouvelles techniques sont fréquemment proposées [TLF08, BMS09].

2.1.3.3 Représentation locale

Après avoir étudié les techniques de détection et de description de POI, nous aborderons le cœur du problème de représentation. Comment extraire une signature pertinente d'un objet représenté par un ensemble de POI ? Il est vrai qu'à chaque POI sont associés des coordonnées et des valeurs d'échelle et d'orientation, mais ces caractéristiques ne sont pas suffisantes pour caractériser la forme de l'objet. De même, s'il est vrai qu'on dispose d'un ensemble de descripteurs extraits autour de chaque POI, ces descripteurs ne peuvent pas être directement utilisables pour caractériser l'apparence de l'objet.
Le principal problème réside dans le fait que les objets sont représentés par un nombre variables de POI, alors que les classifieurs nécessitent un vecteur d'entrée de taille fixe. Ainsi, il convient de faire des statistiques globales sur le nuage de points (ensemble de POI détectés) afin d'extraire des caractéristiques de formes ou de texture. Quant aux caractéristiques d'apparences, il convient d'apparier les descripteurs de POI avec un modèle qui englobe les apparences des objets. Cette reformulation du problème consiste donc à construire un modèle qui combine les représentations locales pour former une représentation de l'ensemble des images d'un objet. Nous en arrivons ainsi au concept de *Vocabulaire Visuel* (codebook, en anglais) que nous allons détailler dans la section suivante.

2.1.3.4 Vocabulaire Visuel

La notion de *vocabulaire* a été utilisée initialement pour la catégorisation de documents écrits. Cela revenait à constituer un vocabulaire de mots afin de repérer les mots utiles pour discriminer des catégories de documents. Par analogie, le même mécanisme peut être invoqué dans le domaine de vision afin de reconnaitre

des catégories d'objet. Ici, on parle de Vocabulaire Visuel (On notera désormais VV le Vocabulaire Visuel) qui englobe un ensemble de mots visuels représentant chacun un groupe de régions similaires.
En pratique, construire ce vocabulaire visuel revient à quantifier, dans un premier temps, l'espace des descripteurs. En second temps, les descripteurs similaires sont regroupés par des méthodes de clustering formant des clusters. Les centroïdes des clusters représentent les mots visuels du vocabulaire et leur nombre représente, à son tour, la taille du vocabulaire.
Une des premières approches utilisant un vocabulaire visuel appris à partir d'un ensemble d'images est celle de Weber et al [WWP00]. La quantification produisant le vocabulaire est réalisée sur les voisinages locaux (des patches) des POI en utilisant l'algorithme de k-means [Mac67]. Depuis, le regroupement de POI (Harris, Harris-Affine, SIFT, etc) avec l'algorithme de k-moyennes est devenu très populaire pour la construction de vocabulaire visuel [SZ03, CDF$^+$04, MS06].
L'algorithme K-moyennes [Mac67] est l'une des méthodes de classification non supervisée les plus simples et les plus utilisées. Etant donnée un ensemble de POI, cet algorithme permet de les partitionner en K groupes de régions similaires par une procédure itérative qui les amène progressivement dans des positions finales stables. Lors des itérations, chaque point est assigné au plus proche des K centres de gravité (centroïdes) des clusters, ensuite les valeurs de centroïdes sont de nouveaux calculées. Cette technique, bien qu'elle permette de réduire la complexité du clustering, converge souvent vers des optima locaux. La principale raison qui motive l'utilisation de cette technique est la faible complexité et consommation en mémoire. En effet, d'un côté, il est impératif d'utiliser une technique rapide quand il s'agit de regrouper des descripteurs de grandes dimensions extraits à partir de centaines d'images. De l'autre côté, le fait de fixer initialement le nombre de clusters risque de ne pas produire de clusters compacts. Il existe d'autres méthodes de clustering fondées sur des algorithmes d'agglomération [DE84], qui déterminent automatiquement le nombre de clusters à travers des regroupements successifs selon un seuil bien déterminé. Dans [LLS04, MLS06, Lei08], un algorithme de clustering agglomératif, appelé RNN (Reciprocal Nearest Neighbor), est utilisé pour construire le vocabulaire visuel. Cet algorithme bien qu'il soit de faible complexité, il converge vers des optima globaux. Le principe repose sur la construction de chaines NN (Nearest-Neighbor chain) constituées par des voisins proches. Dès l'obtention d'une chaine NN, les clusters correspondants sont regroupés et une nouvelle chaine sera construite dans l'itération suivante avec un nouveau point

choisi aléatoirement. Dans [AAR04], une approche similaire de clustering a été utilisée, mais en incorporant des relations géométriques entre les mots du vocabulaire. Jusque là, nous avons traité la problématique de construction d'un modèle d'apparence locale. Il reste à présenter comment extraire un vecteur de caractéristiques en utilisant le vocabulaire visuel afin de caractériser une catégorie d'objet.

2.1.3.5 Extraction de caractéristiques en utilisant le vocabulaire visuel

Vu que la notion de " vocabulaire " a été introduite pour la problématique de catégorisation de documents, il est normal de s'y référer initialement pour appuyer les techniques de caractérisation que nous allons proposer. Le modèle par "sac de mots" est une approche qui s'est avérée très performante dans le domaine de catégorisation de documents [Joa98]. Le principe consiste à représenter chaque document par un histogramme basé sur la fréquence d'apparition de chaque mot du vocabulaire. Par analogie, il est possible d'extraire une signature des images en s'appuyant sur les fréquences d'apparition des régions similaires à celles codées dans le vocabulaire visuel. Ainsi, à chaque image on associe un histogramme dont les bins représentent les mots visuels du vocabulaire et les poids sont les fréquences d'apparition de ces bins dans l'image. Cette représentation, appelée "Bag of features (BoF)", est souple et assez robuste aux variations d'apparence intra-classe. En revanche, ces avantages ont tendances à créer de faux positifs vu que le vocabulaire visuel n'est qu'un modèle génératif qui permet de représenter les apparences. Afin d'améliorer les performances de reconnaissance, l'intégration d'une technique de discrimination, basée sur la classification, s'avère essentielle. Le classifieur le plus utilisé à cet effet est le SVM [GD07, LBH08] qui permet d'avoir recours à des noyaux spécifiques entre histogrammes. Nous revenons sur la technique de classification par SVM dans la section 2.2.2.
L'inconvénient majeur de la caractérisation par histogramme est que ce dernier ne tient pas compte de la disposition spatiale des POI dans l'image. Le fait de ne pas inclure des règles spatiales comme des contraintes géométriques diminue la rigueur d'interprétation de l'image. Agarwal et al. [AAR04] proposent d'intégrer dans le vocabulaire des relations spatiales entre les clusters comme la distance et la direction entre chaque paire. Ensuite, l'image est représentée simplement par un vecteur binaire qui vérifie non seulement l'activation des clusters du Vocabulaire Visuel (on le note désormais VV), mais aussi les relations géométriques entre les descripteurs. Cette approche a été expérimentée pour la détection de véhicules et a obtenu un résultat de détection satisfaisant. Quant à la détection de piétons,

nous ferons largement référence aux travaux de Leibe [LLS04, LSS05, Lei08]. Un modèle implicite de forme (Implicit Shape Model) a été proposé par *Leibe et al.* afin de représenter à la fois l'apparence et la forme des piétons. Ce modèle n'est qu'une variante d'un VV dont la spécifité consiste à associer pour chaque descripteur sa distance par rapport au centre de l'objet. Ainsi, le VV peut être défini comme un modèle de distribution non-paramétrique d'un ensemble de motifs qui caractérisent l'apparence locale de piétons. En ce qui concerne la détection, un système de vote probabiliste a été proposé afin de générer des hypothèses sur les positions des piétons. Quant à la classification, [FLCS05] propose de valider les hypothèses de détection en utilisant un SVM qui opère sur l'activation des clusters du VV tout en intégrant des contraintes d'emplacement des POI.
Nous considérons que le couplage des représentations fondées sur les VV avec des approches discriminatives s'avère nécessaire et permet d'en cumuler les avantages. En effet, la représentation locale, bien qu'elle soit bien adaptée aux larges variations d'apparences et de formes intra-classe, nécessite d'être couplée à une technique de classification permettant de catégoriser précisément les objets routiers. C'est ainsi que la plupart des méthodes utilisant des VV font appel à un système de classification. Dans la section suivante, nous présentons les principaux algorithmes de classification automatique qui ont été utilisés pour la catégorisation d'objets routiers.

2.2 Classification des obstacles routiers

Comme nous l'avons mentionné précédemment, la problématique de la reconnaissance de catégories d'objets se dessine autour des problèmes de représentation et de classification. Après avoir étudié les principales méthodes de représentation, nous faisons le point dans cette section sur les techniques utilisées en littérature pour la classification des OR. Dans le chapitre précédent, nous avons montré que la problématique de classification se situe essentiellement au niveau de la discrimination entre les caractéristiques extraites des images d'OR. Les modèles discriminants cherchent à déterminer une fonction de décision optimale dans l'espace de représentation des vecteurs des caractéristiques.
Les modèles discriminants les plus couramment utilisés sont les machines à vecteurs support (SVM) [PTP98, DT05, SGRB05, ARB09b] et les cascades de classifieurs [VJS03, PSZ08, NCHP08], et d'une façon moins intense les réseaux de neurones [ZT00, SBM06]. Les SVM ont montré leur efficacité pour la discrimination des OR.

Leur avantage majeur est qu'ils s'adaptent facilement aux problèmes non linéairement séparables. En ce qui concerne les modèles de cascade de classifieurs, ils sont facilement transposables pour des applications temps réel. Les détails concernant ces techniques sont exposés dans la suite.

2.2.1 Cascade de classifieurs

Un problème complexe tel que la reconnaissance des OR nécessite un très grand nombre de caractéristiques conduisant à des temps de traitement très importants. Pour réduire la charge de calcul, [JVJS03] ont proposé de construire une cascade de classifieurs de complexité croissante sélectionnant un nombre faible de caractéristiques aux premiers étages.
Le modèle de cascade de classifieurs est essentiellement utilisé pour la détection et la reconnaissance d'OR représentés par régions [TPM06, PSZ08, NCHP08]. Comme nous l'avons expliqué dans la section 2.1.2, l'approche par régions consiste à décomposer l'image d'un objet en nombreuses régions avant de caractériser chacune d'entre elles. Ensuite, des classifieurs peuvent être employés sur ces régions indépendamment les uns des autres avant de fusionner l'ensemble des décisions. Etant donné que plusieurs décisions prises sur des régions peuvent être négatives (pas correctes), il est avantageux de pouvoir rejeter l'objet en question avec le moins possible de calculs. Cela revient à construire une cascade de classifieurs de complexité croissante où les classifieurs les plus simples et les plus rapides sont situés au début. Les cas difficiles ne sont traités que par les derniers classifieurs en s'appuyant généralement sur un nombre important de caractéristiques. Cette technique permet bien évidemment d'obtenir des temps de reconnaissance très courts. La figure 2.8 illustre l'architecture de la cascade.

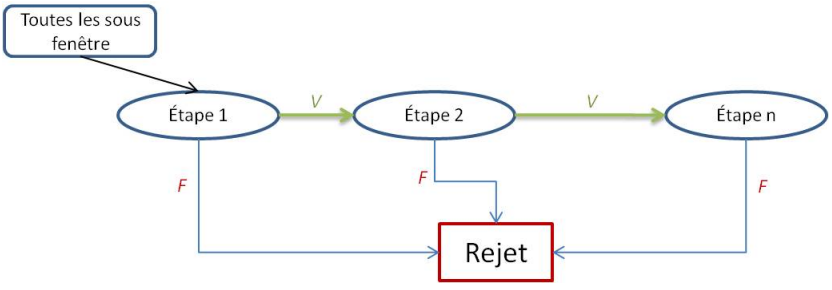


Figure 2.8. Illustration de l'architecture de la cascade de classifieurs

Comme le montre la figure 2.8, la cascade est constituée d'une succession d'étages. En pratique, chacune est formée d'un classifieur fort appris par l'algo-

rithme d'AdaBoost [FS96]. Le principe général de cet algorithme est de combiner un ensemble de classifieurs "faibles" (un peu meilleur que le hasard) en un classifieur fort (performant). En pratique, l'algorithme recherche itérativement dans le vecteur de caractéristiques, les fonctions de classification faibles les plus discriminantes pour les combiner en une fonction de classification forte [NCHP08]. Soit f et h sont les fonctions de classification respectivement forte et faible, et α un coefficient de pondération. La fonction de décision f de la cascade de classifieurs est définie par :

$$f(x) = Signe(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{2.6}$$

Les grandes lignes du pseudo-code de l'algorithme d'adaboost sont données ci dessous (algorithme 2.2.1). Le principe consiste à augmenter à chaque itération les poids des exemples mal classés et de diminuer ceux des éléments bien classés.

Algorithme 1 L'algorithme d'Adaboost

1: Soit une base d'apprentissage contenant N exemples et un nombre maximal d'itérations T
2: Initialiser $\omega_i = \frac{1}{N}, i = 1, ..., N$
3: **Pour** $t = 1, ..., T$ **Faire**
4: Entraîner un classifieur faible h_t à partir des ω_i
5: Calculer l'erreur pondéré ϵ_t
6: Calculer le coefficient de pondération α_t à partir de la valeur de ϵ_t
7: Mettre à jour les poids $\omega_i^{t+1}, \forall i \in [1, N]$
8: **Fin Pour**
9: Sortie : $f(x) = Signe(\sum_{t=1}^{T} \alpha_t h_t(x))$

La reconnaissance d'un objet par cascade de classifieurs est souvent à la fois performante et rapide. En revanche, l'inconvénient majeur de son implémentation est qu'elle requiert un temps d'apprentissage très élevé.

2.2.2 Classification par SVM

Les machines à vecteurs supports (Support Vector Machines, SVM) sont des méthodes généralistes d'apprentissage et de discrimination. Ces méthodes ont montré leur efficacité dans de nombreuses applications, notamment pour la reconnaissance des OR [PTP98, DT05, SGRB05, ARB09b]. Les SVM ont été à l'origine des travaux de Vapnik [Vap95] et ont été conçus pour la décision binaire. L'ori-

ginalité principale de ces méthodes consiste à utiliser efficacement les exemples étiquetés afin de produire une fonction de décision qui maximise la marge entre deux classes données. L'efficacité repose sur la sélection des éléments d'apprentissage les plus représentatifs de la tâche de décision tout en maximisant la capacité de généralisation du modèle. Ces éléments correspondent aux exemples proches de la frontière de décision. Ils sont appelés *les supports vecteurs* et sont situés sur la marge (voir figure 2.9). Pour garantir la bonne généralisation de la fonction de décision recherchée, le problème mathématique correspond à la maximisation de la distance des exemples annotés de la frontière.

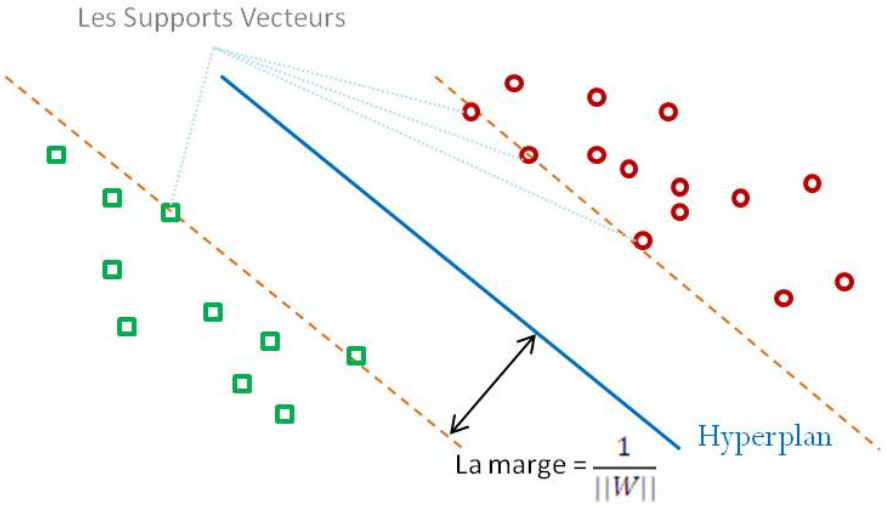


Figure 2.9. Séparateur à vaste marge

Soit $X \in \mathbb{R}^{n \times d}$ une liste de n exemples annotés de vecteurs caractéristiques $x \in \mathbb{R}^d, i \in [1, n]$. Les $y_i \in \{-1, 1\}, i \in [1, n]$, sont les étiquettes binaires associées. Tout élément x situé sur la frontière de décision vérifie l'équation $(f(x) + b = 0)$, avec $b \in \mathbb{R}$ le biais. Dans le cas où les données sont séparables [3], le problème mathématique s'écrit :

$$\begin{cases} \min\limits_{f,b} & \|f\| \\ s.c. & y_i(f(x_i) + b) \geq 1, i \in [1, n] \end{cases}$$

Dans le cas des classes non séparables, des variables de relâchement $\xi_i \geq 0$ sont introduites afin de considérer les erreurs opérées pendant la phase d'apprentissage. Formulé ainsi, le problème se ramène à :

3. sans erreurs de classification par un hyperplan linéaire

$$\begin{cases} \min\limits_{f,b,\xi_i,i\in[1,n]} & \frac{1}{2}\|f\|^2 + C\sum\limits_{i=1}^{n}\xi_i \\ s.c. & y_i(f(x_i)+b) \geq 1-\xi_i, i\in[1,n] \\ & \xi_i \geq 0, i\in[1,n] \end{cases}$$

Avec $C \in \mathbb{R}$ est un paramètre utilisé pour quantifier l'importance du relâchement des sous contraintes (s.c). Ensuite, l'utilisation de la méthode du Lagrangien [4] mène à une formulation duale du problème :

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum\limits_{i=1}^{n}\alpha_i \\ s.c. & \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, i\in[1,n] \end{cases}$$

Avec $\alpha = \alpha_1, ..., \alpha_n$ sont les multiplicateurs de Lagrange. Le lecteur pourra se référer à [Sua06] où l'ensemble des calculs intermédiaires est détaillé. La fonction de décision est définie finalement par :

$$\begin{aligned} f(.) &= \sum_{i=1}^{n}\alpha_i y_i K(x_i, .) \\ dec(x) &= Signe(f(x)+b) \end{aligned} \tag{2.7}$$

Avec $K(\cdot,\cdot)$ est un noyau symétrique et défini positif qui permet d'évaluer la similarité entre deux vecteurs x et x'. Nous aborderons les principales formes que peut prendre la fonction noyau dans la section suivante.

2.2.2.1 Les hyperparamètres de SVM

La mise en œuvre de SVM requiert la détermination des valeurs d'hyperparamètres. Bien que ces paramètres soient inconnus a priori, ils sont décisifs pour obtenir un modèle performant de classification. Les deux principaux paramètres à régler sont l'influence du noyau et de ses paramètres, et la complexité du modèle contrôlée par le paramètre de pénalité C.
Un noyau $K(.,.) : (\mathbb{R}^d)^2 \mapsto \mathbb{R}$ permet de mesurer la similarité entre deux vecteurs caractéristiques x et x'. Le tableau 2.2 illustre les deux noyaux les plus utilisés ainsi que leurs paramètres spécifiques.

4. une méthode usuelle de résolution de problèmes sous contraintes

Table 2.2. Les noyaux couramment utilisés pour comparer des vecteurs de caractéristiques

Nom	Paramètre	Linéarité	Formule
Polynomial	Ordre : $p \geq 1$	Linéaire pour $p = 1$	$K(x, x') = (< x, x' > +1)^p$
Gaussien (RBF)	Largeur de bande : σ	Non-linéaire	$K(x, x') = \exp(-\frac{1}{2}\frac{\|x-x'\|^2}{\sigma^2})$

Les valeurs des hyperparamètres du modèle ont des effets croisés. En effet, les machines à noyaux peuvent donner à la fois les meilleurs et les moins bons résultats pour des valeurs différentes d'hyperparamètres. La manière classique de les régler est de chercher sur une grille, les valeurs optimales au sens d'un critère de validation [Can07] comme la précision, le rappel ou l'erreur empirique. Par exemple, pour un noyau RBF, les deux paramètres principaux à régler sont la valeur du C et la largeur de bande σ. Le meilleur couple du paramètre est déterminé par validation en le cherchant sur une grille de deux dimensions.

2.2.2.2 SVM pour la décision multiclasse

Les méthodes classiques d'utilisation des SVM pour le cas multiclasse consistent à décomposer, dans un premier temps, le problème en une série de dichotomies (un-contre-un, un-contre-tous). Ensuite, les décisions des classifieurs élémentaires sont combinés par une stratégie de fusion (vote, utilisation de la théorie probabiliste ou de l'évidence) pour permettre la discrimination multiclasse. L'approche un-contre-un propose d'utiliser $\frac{n(n-1)}{2}$ discriminateurs binaires pour décrire toutes les dichotomies possibles parmi les n classes. Quant à l'approche "un-contre-tous", elle utilise n classifieurs binaires dont chacun est spécialisé pour la reconnaissance d'une classe opposée à la fusion des $(n-1)$ autres classes. Du point de vue performance de classification, aucune de ces approches n'est meilleure dans tous les cas. Toutefois, il convient d'utiliser l'approche un-contre-un afin de pouvoir associer des probabilités à chacun des SVM binaires. L'application d'une méthode de fusion à partir de ces probabilités semble donner de meilleurs résultats [HT98, Can07]. Concernant ce dernier point, une étude détaillée des méthodes de fusion est présentée dans la section suivante.

2.3 Fusion d'informations pour la reconnaissance des obstacles routiers

Plusieurs études ont montré l'apport de la fusion des données, issues de différents capteurs [PLR+06, BBFV06, BBG+07, FC08, ARB09b, ARB09a]. Notre choix, comme nous l'avons justifié dans la section 1, s'est porté sur le traitement de données issues de capteurs de vision. Un système intéressant, appelé Tetravision, a été proposé dans [BBFV06] et basé sur la combinaison d'une paire de systèmes stéréos VIS et IR. Initialement, les deux caméras stéréo sont traitées indépendamment, puis les résultats de détection de piétons sont combinés au moyen d'un "OU logique". Dans d'autres systèmes [ARB09a], la fusion se produit au niveau de classifieurs spécialisés pour chaque modalité. Dans ces cas, il s'agit de combiner les décisions établies par différents capteurs ou classifieurs pour prendre une décision fiable. Ce niveau de fusion est appelé, haut niveau et les approches principales de fusion employées sont les approches par vote ou par utilisation des théories probabilistes ou de l'évidence. À un stade plus précoce, la fusion peut intervenir afin de fiabiliser les données, les compléter et les présenter de façon facilement exploitable. Par exemple, dans [ARB09b], un système de reconnaissance d'OR, basé sur la combinaison des caractéristiques extraites du spectre VIS et IR avant la classification, a été proposé. Ces deux niveaux différents de fusion (bas, haut niveau) de données illustrent la typologie de la fusion d'informations.

2.3.1 Fusion de caractéristiques

La fusion au niveau de l'espace des caractéristiques consiste à concaténer, avant l'étape d'apprentissage, toutes les caractéristiques en un seul vecteur. Ce vecteur est ensuite fourni en entrée d'un classifieur.
L'enjeu de la fusion à bas niveau repose sur le choix de la méthode utilisée pour fusionner les caractéristiques. La plus simple méthode consiste à concaténer les vecteurs unimodaux (figure 2.10). Une étape de normalisation des vecteurs unimodaux est souvent requise afin d'éviter que des composantes influent plus que d'autres pour la classification. Plusieurs méthodes de normalisation existent, les deux plus populaires consistent à effectuer une transformation, soit linéaire dans un intervalle prédéfini, soit gaussienne pour que les données suivent une distribution normale centrée réduite (de moyenne 0 et variance 1).

La fusion de caractéristiques a l'avantage de ne nécessiter qu'une seule phase d'apprentissage automatique pour l'ensemble des modalités. Ceci permettra au

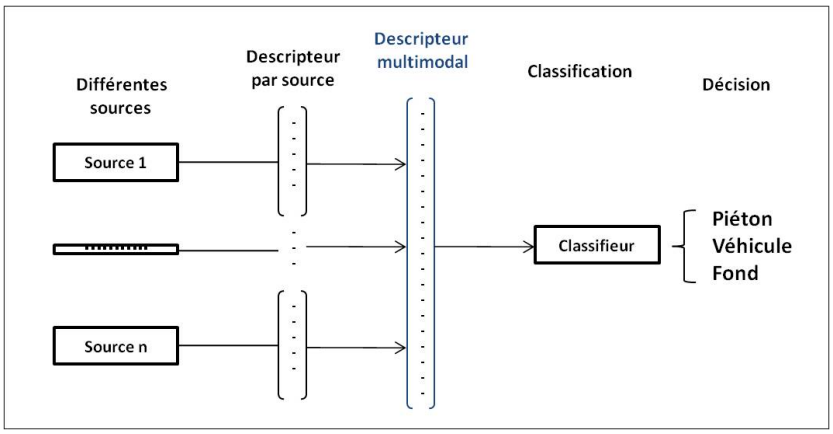


Figure 2.10. Description générale d'un processus de fusion de caractéristiques

classifieur d'apprendre des régularités dans un espace multimodal [Aya07]. En revanche, la fusion de bas niveau induit l'utilisation d'une méthode de classification unique pour tous les types de descripteurs. Pour certains algorithmes de classification, tels que les SVM, il peut être intéressant d'avoir recours à des noyaux différents selon les modalités.
Dans certains cas, la concaténation des vecteurs caractéristiques peut conduire à un espace de grande dimension dans lequel la phase d'apprentissage peut ne pas converger. Pour remédier à ce problème, il convient de réduire le nombre de caractéristiques, typiquement en appliquant un algorithme de sélection d'attributs.
La dimension d'un vecteur caractéristique a une très forte influence sur les performances des systèmes de classification automatique. En effet, Il est difficile de s'assurer du bon fonctionnement, même d'un bon algorithme d'apprentissage, quand l'information est représentée par un grand nombre d'attributs non pertinents. L'objectif de la sélection de caractéristiques est de trouver un sous ensemble optimal constitué d'attributs pertinents qui permet d'éliminer les attributs redondants tout en conservant la précision.
La majorité des méthodes de sélection d'attributs passent par quatre étapes, comme le montre le schéma général présenté par [DL97] (figure 2.11).

Génération de sous ensembles d'attributs

Pour former un sous ensemble d'attributs, il faut partir d'un ensemble de départ et opter pour une stratégie de recherche. L'ensemble de départ peut être l'ensemble de tous les attributs disponibles, un ensemble vide ou un ensemble d'attributs tirés de manière aléatoire. Si n présente le nombre d'attributs disponibles,

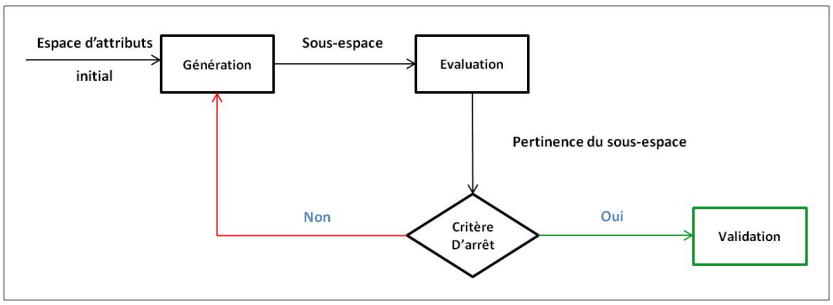


Figure 2.11. Illustration de la procédure de sélection d'attributs

l'espace de recherche comportera 2^n sous ensembles candidats. Plusieurs stratégies de recherche heuristiques peuvent être envisagées. Elles peuvent être classées en trois catégories : la recherche complète, la recherche séquentielle et la recherche aléatoire. Une description détaillée de ces stratégies est donnée dans [Por09].

Stratégie d'évaluation

Ce processus mesure la pertinence du sous ensemble généré suite à la procédure de génération. Parmi les mesures d'évaluations utilisées, nous citons les mesures de séparabilité [Fuk90], dépendance [DL97], corrélation [Hal98], erreur de classification [GE03], consistance [Sem04] et d'information [MCB06].

Critère d'arrêt

Lorsque le critère d'arrêt est satisfait, la procédure de la recherche des sous espaces d'attributs s'arrête. Les critères d'arrêt les plus fréquents sont liés à une valeur seuil (nombre minimum d'attributs, un nombre maximale d'itérations ou taux d'erreur de classification). La recherche est donc non exhaustive, mais réalisable dans la pratique.

Validation des résultats

Ce processus consiste à valider le sous espace d'attributs sélectionné en comparant, généralement, les résultats de classification.

Bien évidemment, un algorithme de sélection d'attributs permet de réduire ensuite les temps d'apprentissage et d'exécution ce qui rend également le processus d'apprentissage moins coûteux. Cependant, la fusion de caractéristiques, suivie par une sélection des attributs les plus pertinents, est limitée dans ses usages. Elle suppose en effet que les caractéristiques soient de nature similaire. Une solution,

plus complexe, mais plus souple et appropriée consiste à utiliser un classifieur pour chaque catégorie de caractéristiques et à fusionner ensuite les décisions des classifieurs.

2.3.2 Fusion de décisions

Par opposition à la fusion des caractéristiques, la fusion de classifieurs (figure 2.12), consiste à fusionner les décisions prises séparément pour chaque système monomodal. Ce type de fusion est appelé fusion haut niveau car la fusion tient en compte les décisions établies par les différents classifieurs afin de déterminer la catégorie de l'objet à classifier.

2.3.2.1 Les motivations d'utilisation

L'idée principale derrière la combinaison de classifieurs est l'amélioration de la prise de décision. En effet, il n'existe pas une méthode de décision qui arrive à satisfaire entièrement les exigences d'un problème. La combinaison de plusieurs décisions, permet éventuellement d'en cumuler les avantages. Ainsi, la combinaison de classifieurs est considérée comme une excellente alternative à l'utilisation d'un unique classifieur.

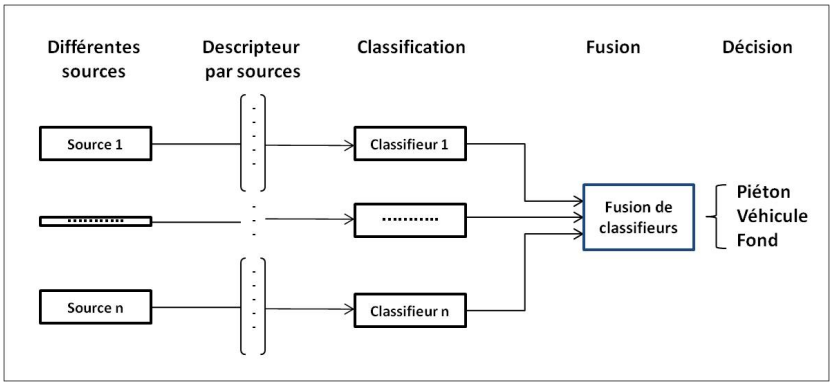


Figure 2.12. Description générale du processus de fusion de classifieurs

On étudie ici la fusion des classifieurs comme étant un problème de fusion des résultats de classification (voir figure 2.12). Étant donné que les informations fournies sont *incertaines* ou *imprécises*, elles ne permettent pas de classifier avec certitude un objet. Cet aspect est étudié dans la section suivante.

2.3.2.2 Théorie de l'incertain pour la fusion de données

Il est important de prendre en compte les imprécisions et les incertitudes des informations à combiner. Une proposition peut être imprécise, incertaine ou à la fois imprécise et incertaine. L'incertitude caractérise un degré de conformité à la réalité (défaut qualitatif de l'information), tandis que l'imprécision mesure un défaut quantitatif de l'information (par exemple une erreur de mesure) [LM09].
Les théories de l'incertain représentent les imperfections des informations par l'intermédiaire d'une fonction de mesure de confiance. D'une façon générale, l'application de la théorie de l'incertain dans un processus de fusion des décisions de classifieurs est soumise à trois étapes principales qui sont illustrées dans la figure 2.13. La première étape consiste à modéliser les connaissances en attribuant une valeur de confiance $M_j(C_i)$ à chaque classe C_i. Cette valeur de confiance est concrètement mise en œuvre à travers la fonction de décision de SVM ($f(x_i)$). Avec x_i représente le vecteur caractéristique extrait à travers l'analyse des données provenant de la source j. Après la modélisation des connaissances, il s'ensuit une étape de combinaison des mesures de confiances par une règle de combinaison. Finalement, la troisième étape consiste à prendre une décision, selon un critère bien défini, sur l'attribution de l'objet en question à une classe C.

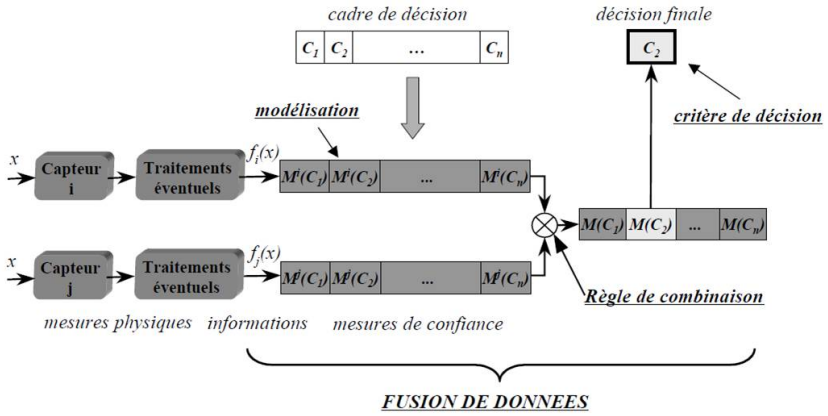


Figure 2.13. Illustration des trois étapes de fusion dans le cadre de la combinaison des décisions de classification

Parmi les modèles les plus couramment utilisés en fusion de données, on cite l'approche bayésienne, la théorie des probabilités et la théorie des croyances. Le choix d'un modèle de fusion est souvent conditionné par la facilité de sa mise en œuvre et surtout son aptitude à modéliser des connaissances de nature incertaine et/ou imprécise. Les performances du modèle bayésien de fusion sont largement

dépendantes de la précision de l'estimation des distributions de probabilité. La théorie des probabilités est de plus largement employée pour des problèmes d'estimation [FM09]. Les méthodes classiques issues de cette théorie consistent à utiliser des opérateurs de combinaison de probabilités tels que le maximum, la somme, le produit, etc. Ces méthodes estiment des probabilités a posteriori des différentes classes en supposant que les classifieurs sont indépendants. Ainsi, la règle de décision consiste à sélectionner la classe $C_i, i \in [1, n]$ pour laquelle la probabilité a posteriori P_i est la plus élevée. Il est possible de pondérer les probabilités en fonction de la fiabilité des classifieurs. Soit Q désignant le nombre de sources à combiner, le tableau 2.3 définit les principaux opérateurs de fusion probabiliste, à savoir la somme, le produit, avec et sans pondération.

Table 2.3. Les principaux opérateurs de fusion

Règle de combinaison	Formule
Somme	$P_i = \sum_{j=1}^{Q} p_{i,j}$
Produit	$P_i = \prod_{j=1}^{Q} p_{i,j}$
Somme pondérée	$P_i = \sum_{j=1}^{Q} \omega_j p_{i,j}$
Produit pondéré	$P_i = \prod_{j=1}^{Q} \omega_j p_{i,j}$

La théorie des croyances est souvent considérée comme une généralisation de la théorie des probabilités. Elle est basée sur un fondement mathématique robuste qui permet de représenter et de manipuler des informations entachées d'incertitude et d'imprécision.
La théorie des croyances est introduite par Shafer [Sha78] et plus tard reprise par Smets [SK94] dans son modèle des croyances transférables (MCT), elle porte également le nom de théorie de Dempster-Shafer (DST) ou théorie de l'évidence. Elle est largement employée dans le cadre de la fusion d'informations pour améliorer l'analyse et l'interprétation des données issues de sources d'informations multiples. Le MCT est basé sur la définition de fonctions de croyance fournies par des sources d'information pouvant être complémentaires, redondantes et éventuellement non indépendantes. Les mécanismes de raisonnement du DST peuvent être regroupés en deux niveaux : le niveau crédal et pignistique. Comme illustré dans la figure 2.14, le niveau crédal permet la représentation et la manipulation des croyances, tandis que, le niveau pignistique, est utilisé pour la prise de décision dans un cadre probabiliste.

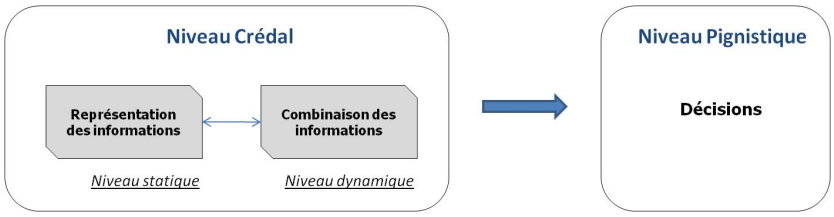


Figure 2.14. Représentation abstraite des mécanismes en MCT

Nous revenons en détail sur le formalisme des fonctions de croyance et son application dans la section 4.2 .

2.4 Bilan

Pour répondre à la problématique de la reconnaissance de catégories d'objets dans les images, nous avons axé notre réflexion autour de trois axes : la représentation, la classification et la fusion d'informations. Après avoir dressé l'état de l'art, nous faisons le bilan afin de pouvoir synthétiser les différentes méthodes et orienter nos choix.

Les problèmes récurrents à résoudre résident dans les larges variations de formes et d'apparences, ainsi qu'aux problèmes d'occultations partielles d'objets. Ainsi, il faut donc mettre au point des méthodes de représentation et de classification suffisamment robustes.

Les deux critères les plus importants à considérer dans le choix de la méthode de représentation est le pouvoir de généralisation et la pertinence. Le pouvoir de généralisation se traduit par le fait de pouvoir représenter une très grande variété d'une catégorie d'objets à l'aide d'un faible échantillon en apprentissage. Parmi les méthodes de représentation proposées, la caractérisation locale semble la plus appropriée au regard de ce critère. De plus, cette approche permet de surmonter les problèmes d'occultations partielles en caractérisant les objets à l'aide de descripteurs locaux. Face à la diversité d'apparence des obstacles routiers, notamment les piétons, les caractéristiques globales extraites sont à éviter. Toutefois, la combinaison de ces caractéristiques avec des caractéristiques locales peut apporter des informations complémentaires et amener à une caractérisation pertinente.

Le pouvoir de généralisation de la méthode de représentation peut être renforcé par la définition d'une base d'apprentissage contenant une grande variété d'apparences et de formes pour chaque catégorie. Nous proposons ensuite, d'utiliser un classifieur de type SVM, permettant de ne retenir que les exemples les plus

discriminants pour la classification. Ce choix est justifié non seulement, car SVM est réputé pour traiter des données de grande dimension, mais aussi parcequ'il a montré son efficacité pour la discrimination non linéaire des OR. Ayant justifié l'apport du processus de fusion d'informations, nous proposons également d'utiliser des techniques de fusion de données afin d'améliorer l'analyse et l'interprétation des données.

2.5 Conclusion

Dans ce chapitre consacré à l'état de l'art, nous avons présenté les techniques de base nécessaires à la reconnaissance de catégories d'objets dans les images. Notre étude s'est focalisée plus particulièrement sur les étapes de représentation, de classification et de fusion d'informations. Les différentes méthodes de caractérisation proposées en littérature ont été divisées en trois catégories : globale, par région et locale. Concernant la classification, nous avons présenté les deux modèles de discrimination les plus utilisés : SVM et la cascade de classifieurs. Finalement, nous avons illustré les techniques et les mécanismes de fusion de données en étudiant d'une façon plus approfondie la théorie des fonctions de croyance. Dans cette dernière partie, nous avons montré que le processus de fusion constitue un moyen intéressant pour l'amélioration des performances d'un système de reconnaissance d'obstacles routiers.
L'examen de l'état de l'art nous a déjà permis d'écarter quelques méthodes inadaptées à nos besoins. Les choix opérés et les méthodes mises en oeuvre sont exposés dans le chapitre suivant.

Chapitre 3

Vocabulaire Visuel Hiérarchique pour la catégorisation des obstacles routiers

Introduction

Dans le chapitre précédent, nous avons recensé à travers notre état de l'art les différentes méthodes de représentation et de classification des obstacles routiers. Dans ce chapitre, nous présentons notre contribution qui se décline selon deux axes : 1) La conception d'un modèle d'apparence locale basé sur un ensemble de descripteurs représentés dans un Vocabulaire Visuel Hiérarchique. 2) La combinaison du modèle d'apparence avec un classifieur SVM. Puisque le Vocabulaire Visuel ne permet de représenter que les apparences locales des objets, nous proposons également l'intégration des caractéristiques globales extraites à partir des points d'intérêts robustes à la translation, à l'échelle et à la rotation.
À la fin du chapitre, nous présentons les résultats obtenus suite à l'expérimentation de notre méthode sur deux bases d'objets routiers.

3.1 Vers un modèle de représentation locale et globale

Après avoir présenté la problématique (chapitre 1) et l'état de l'art (chapitre 2) sur les méthodes de représentation, nous justifions dans cette section, le choix du modèle de représentation proposé. Ce modèle se base sur la combinaison de caractéristiques locales et globales. Extraites à partir d'un ensemble de POI, les ca-

ractéristiques locales permettent de caractériser les apparences locales des objets. Comme nous l'avons expliqué dans la section 2.1.3, cette représentation permet de résoudre les problèmes posés par les larges variations de formes et d'apparences et aussi les occultations partielles de l'objet. Afin de renforcer la pertinence de la représentation, nous proposons d'ajouter des caractéristiques globales extraites à partir de l'ensemble de POI. La caractérisation globale permettra d'apporter des informations complémentaires en caractérisant les formes et les textures globales des objets. Nous expliquons dans ce qui suit, après la justification du choix de POI et du descripteur, les méthodes que nous proposons afin d'extraire ces caractéristiques visuelles.

3.1.1 Choix du point d'intérêt et du descripteur

Dans la littérature, il a été démontré que les détecteurs SIFT et SURF sont actuellement les détecteurs de points d'intérêt les plus utilisés. Ainsi, les auteurs ont accordé une grande importance à ces deux détecteurs en comparant leurs performances avant toute contribution [BTG06, BP07, ETLF11]. Dans notre analyse, nous ne nous appuyons pas seulement sur ces travaux, mais aussi sur une analyse spécifique à notre contexte d'utilisation, à savoir la détection d'OR à partir des images IR.

Rappelons que notre problématique consiste à détecter un obstacle qui se déplace face à un véhicule en mouvement à l'aide d'une caméra embarquée. Cet obstacle est donc sujet à des changements de point de vue, d'échelle et d'illumination. Les résultats comparatifs présentés dans quelques travaux ont montré que le SIFT présente une meilleure robustesse face à la rotation par rapport au SURF. Ce critère est moins important dans notre cas, car la caméra embraquée ne subit qu'une légère rotation. Mais ce qui est plus important c'est, d'une part la rapidité et la robustesse face aux changements d'illuminations, et d'autre part, le fait que le POI soit bien adapté aux problématiques de détection et de reconnaissance d'objets dans les images IR.

Dans [BTG06], l'auteur montre que le SURF surpasse le SIFT en termes de rapidité et de robustesse face aux différentes transformations d'images. En effet, ce détecteur prend les points forts des meilleurs détecteurs et descripteurs l'ayant précédé par :

- le calcul de l'image intégrale et de la convolution,
- l'utilisation des mesures à base de matrices hessiennes rapides,
- la possibilité de description de régions d'intérêt par seulement 64 descrip-

teurs.

Mise à part sa rapidité et sa robustesse, le SURF est bien adapté à la problématique de détection d'OR dans les images infrarouges. En effet, il s'appuie sur la valeur du Laplacien pour indexer des zones d'intérêt. Ce critère permet non seulement d'accélérer le processus de mise en correspondance, mais aussi d'extraire des régions claires dans l'image. Ces régions, ayant une valeur négative du Laplacien, sont à fort contraste et sont caractéristiques de la présence d'objets chauds dans les images IR. Cela permettra évidemment de repérer des zones caractéristiques de présence de piétons ou de véhicules en mouvement.
Comme nous l'avons déjà mentionné dans le chapitre précédent, le SURF et le SIFT sont aussi bien des détecteurs que des descripteurs de POI. Pour décrire une zone d'intérêt, le SIFT se base sur l'extraction des caractéristiques HOG (pour plus de détail, voir section 2.1.2.1). Ces caractéristiques se basent sur la texture et sont plus adaptées au traitement d'images visibles. Quant aux descripteurs SURF, leur calcul se base sur la variation d'intensité entre des régions, ce qui semble plus pertinent pour caractériser des zones d'intérêt dans les images IR.
Pour toutes ces raisons, nous avons choisi d'utiliser l'algorithme SURF-64 afin de détecter et de décrire rapidement les zones d'intérêt. Les POI SIFT et les descripteurs SURF-128 serviront de comparatif. Dans la figure 3.1, nous mettons en valeur quelques POI SURF extraits à partir des imagettes de piétons et de véhicules.

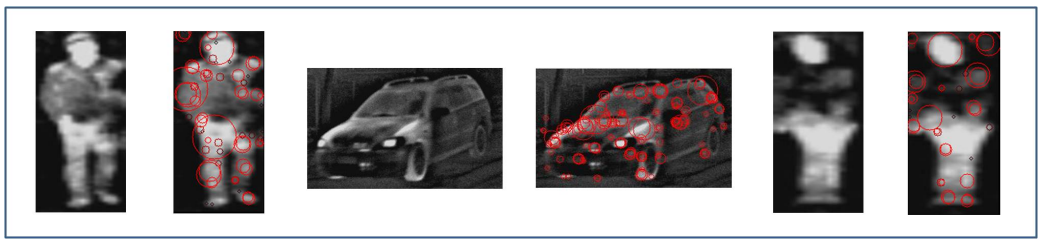


Figure 3.1. Exemples de POI SURF extraits dans des imagettes de piétons et de véhicules. Les cercles en rouge sont dessinés autour des centres de POI ; Le rayon de chaque cercle correspond à la valeur d'échelle à partir de laquelle le point a été extrait.

3.1.2 Extraction des caractéristiques locales

Avant de décrire notre méthode, nous rappelons le principe de la représentation locale. À l'opposé des méthodes globales, les méthodes locales ne considèrent pas les images comme un ensemble de pixels, mais comme une collection de régions locales. Ainsi, la représentation locale consiste à extraire une signature à partir de l'ensemble de descripteurs de POI caractérisant ces régions locales. Cela passera

généralement par la mise en correspondance avec des descripteurs stockés dans un Vocabulaire Visuel qui compacte l'ensemble des apparences locales d'une catégorie d'objet. Lors de la reconnaissance, les descripteurs sélectionnés dans l'imagette sont recherchés dans le VV, ce qui permet de voter pour les imagettes de références qui contiennent des descripteurs similaires. Ainsi, le vecteur caractérisant l'imagette est extrait à partir d'un histogramme de fréquence d'apparition de descripteurs similaires dans le VV.
Une première difficulté se présente dans la définition de la mesure de similarité. La majorité des méthodes évaluent cette mesure en la comparant avec le seuil de clustering utilisé pour définir les mots visuels (clusters) du VV. Typiquement, le seuil est choisi de façon à réduire au maximum la distance entre les descripteurs d'un même cluster tout en augmentant au maximum la distance entre clusters. Le choix de la mesure de similarité est très important. Malheureusement, trop souvent, il s'agit d'un choix arbitraire utilisé pour comparer tous les descripteurs de la même manière. Une nouvelle difficulté apparait lorsque les clusters ont des tailles très variables, présentant ainsi des degrés d'importance différents. Enfin, nous considérons que la mise en correspondance de tous les POI avec des centaines de clusters est une tâche très lente à mettre en œuvre.
Pour pallier ces problèmes, nous proposons une structure hiérarchique pour le VV permettant à la fois de gérer plusieurs niveaux de clustering et d'accélérer les temps de mise en correspondance [BRB10, BLRB10]. Ainsi, la représentation locale d'une imagette est réalisée en trois étapes :

1. Extraction de POI SURF.
2. Mise en correspondance des descripteurs SURF-64 avec le Vocabulaire Visuel Hiérarchique (On notera désormais VVH un Vocabulaire Visuel Hiérarchique).
3. Calcul d'un histogramme de fréquence d'apparition de descripteurs similaires dans le VVH.

Les méthodes proposées pour la construction du VVH et l'extraction des caractéristiques locales, sont présentées dans la section 3.2.

3.1.3 Extraction des caractéristiques globales

L'inconvénient majeur de la caractérisation locale est qu'aucune information sur la disposition spatiale des POI n'est considérée. Le fait de ne pas inclure des règles spatiales ou des contraintes globales diminue la rigueur d'interprétation de

l'image. Ainsi, nous proposons d'augmenter la description vectorielle des imagettes par l'utilisation de descripteurs globaux. Ces descripteurs fournissent des mesures de forme et de texture permettant ainsi d'apporter des informations complémentaires.
À part le descripteur basé sur des sommes de réponses d'ondelettes de Haar, il est associé à chaque POI SURF :

- Une valeur d'échelle (ρ),
- Une valeur de Hessien qui estime la puissance de la réponse de la région autour du POI,
- Une valeur de Laplacien variant du positif au négatif selon le contexte clair ou sombre dans l'image.

Nous proposons ainsi d'exploiter ces caractéristiques pour calculer des statistiques globales sur le nuage de POI extraits de l'imagette. Nous distinguons entre deux ensembles de POI selon le contexte claire ou sombre dans l'image. Dans la figure 3.2, nous mettons en évidence des POI extraits à partir des régions claires et sombres. Pour chacun des deux ensembles, nous déterminons les caractéristiques globales suivantes :

- Rapport $\frac{largeur}{hauteur}$ de la fenêtre (F) qui englobe l'ensemble de POI ;
- Nombre total de points SURF (normalisé par la résolution de F) ;
- Emplacement du centre de gravité des POI ;
- Histogramme de fréquence des directions de contours (8 directions) ;
- Moyenne et écart type des échelles ;
- Moyenne et écart type des valeurs hessiennes.

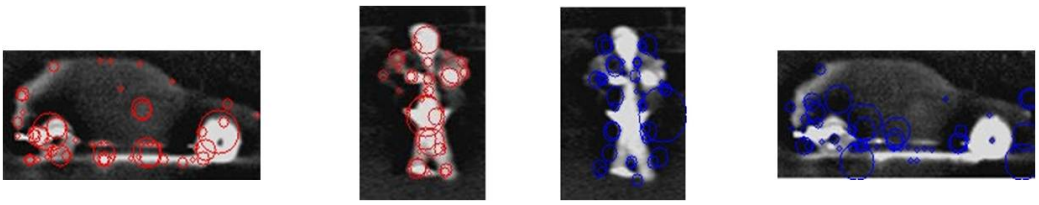


Figure 3.2. Ensemble de POI SURF extraits à partir des régions claires (cercles en rouge) et des régions sombres (cercles en bleu)

Notons que les trois permiers attributs permettent de décrire globalement la forme de l'objet. Les autres permettent, quant à eux, de caractériser la texture. En outre, nous proposons d'enrichir la description globale en ajoutant des caractéristiques de texture calculées sur des pixels comme la moyenne, la variance, le biais et le kurtosis des intensités lumineuses. D'où l'obtention d'un vecteur de 34

caractéristiques globales.

3.2 Représentation des apparences locales dans un Vocabulaire Visuel Hiérarchique

Dans cette section, nous abordons la conception d'un modèle d'apparence locale basé sur un ensemble de descripteurs SURF représentés dans un Vocabulaire Visuel Hiérarchique. La structure hiérarchique a été conçue afin de gérer plusieurs niveaux de clustering et d'accélérer les temps de mise en correspondance avec les descripteurs contenus dans le VV.
Le processus de construction du vocabulaire peut être subdivisé en deux étapes principales :

1. Extraction d'un ensemble de descripteurs d'une base d'images d'apprentissage.
2. Regroupement des descripteurs similaires par des méthodes de clustering.

La figure 3.3 illustre les processus d'extraction et de clustering des descripteurs denses de POI.

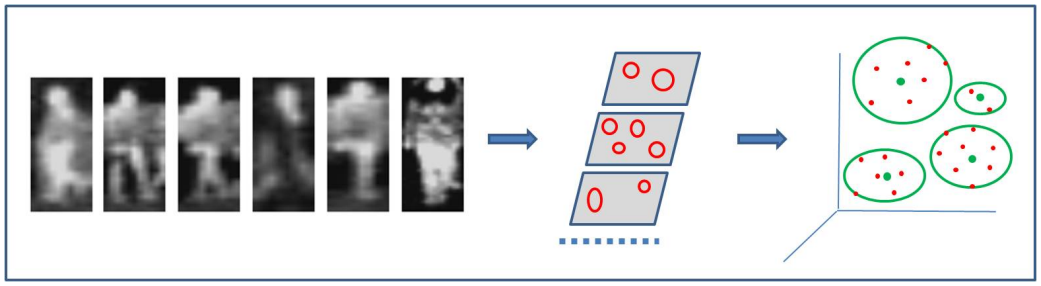


Figure 3.3. Extraction et clustering de descripteurs de POI extraits à partir d'un ensemble de données d'apprentissage. Les cercles en rouge (vert) représentent des POI (des clusters). Les points situés dans le centre des cercles en vert (clusters résultant du processus de clustering) représentent les centroïdes de clusters.

3.2.1 Construction du Vocabulaire Visuel

Afin de construire le VV, nous avons utilisé l'algorithme de clustering RNN (Reciprocal Nearest Neighbor), comme décrit dans [Lei08]. Cet algorithme bien qu'il soit de faible complexité, il converge vers des optima globaux. Le principe repose sur la construction de chaines NN (Nearest-Neighbor chain) constituées par des voisins proches. Dès l'obtention d'une chaine NN, les clusters correspondants sont regroupés et une nouvelle chaine est commencée dans l'itération suivante avec

un nouveau élément choisi aléatoirement. Les détails de l'algorithme sont donnés ci-dessous.

Algorithme 2 Clustering agglomératif en utilisant l'algorithme de RNN

Commencer une chaine C_1 avec un descripteur aléatoire $d \in D$
Inclure tous les descripteurs extraits de la base d'apprentissage dans C_2
$dern \leftarrow 0; dernSim[0] \leftarrow 0; C_1[dern] \leftarrow d \in D; C_2 \leftarrow D \backslash d$
Tantque $C_2 \neq \emptyset$ **Faire**
  $(s, sim) \leftarrow plusProcheVoisin(C_1[dern], C_2)$
  Si $sim > dernSim[dern]$ **Alors**
    $dern \leftarrow dern + 1; C_1[dern] \leftarrow s; C_2 \leftarrow C_2 \backslash \{s\}; dernSim[dern] \leftarrow sim$
  Sinon
    Si $dernSim[dern] > t$ **Alors**
      $s \leftarrow regroupement(C_1[dern], C_1[dern-1])$
      $C_2 \leftarrow C_2 \cup \{s\}$
      $dern \leftarrow dern - 2$
    Sinon
      $dern \leftarrow -1$
    Fin Si
  Fin Si
  Si $dern < 0$ **Alors**
    Initialiser une nouvelle chaine avec un autre descripteur aléatoire $d \in C_2$
    $dern \leftarrow dern + 1$
    $C_1[dern] \leftarrow d \in C_2; C_2 \leftarrow C_2 \backslash \{d\}$
  Fin Si
Fin Tantque

La pertinence du VV construit avec cet algorithme est très dépendante de la valeur du seuil de clustering t. Cela révèle l'importance du choix de la valeur de ce seuil. Dans la section suivante, une structure hiérarchique proposant plusieurs seuils de clustering, plus flexible et plus pertinente, est proposée.

3.2.2 Conception de la structure hiérarchique

Dans cette section, nous proposons une structure hiérarchique pour le VV afin de fournir une représentation plus riche et plus pertinente des apparences locales. En outre, nous associons une fonction d'évaluation permettant d'évaluer la pertinence de la structure.

Le VVH est représenté comme un arbre n-aire [1] dont chaque élément représente un cluster caractérisé par :

- un centroïde C_i : vecteur descripteur moyen calculé sur les descripteurs appartenant au cluster ;
- un rayon R_i : distance Euclidienne entre le centroïde et le descripteur les plus éloigné dans le cluster ;
- une taille n_i : le nombre de points d'intérêt contenus dans le cluster.

Chaque niveau hiérarchique de l'arbre (l) résulte de l'application de l'algorithme de clustering agglomératif RNN avec un seuil spécifique. Le rayon des clusters dans le niveau le plus bas (feuilles) est inférieur au seuil initial ${t_1}^\theta$, alors que le rayon du cluster correspondant à la racine est inférieur au seuil t_{max}. La valeur ${t_{max}}^\theta$ est définie comme étant la plus petite valeur de seuil permettant de grouper tous les descripteurs SURF dans un même cluster (racine de l'arbre).
${t_1}^\theta$ et l^θ (profondeur optimale) sont obtenus (dans l'ordre) en minimisant une fonction d'évaluation $F(t;l)$. Cette fonction implique le taux de clusters unitaires et le taux maximal de points contenus dans les clusters (Eq 3.1).

Soit N désigne le nombre total de clusters dans le VVH.
N inclut N_u clusters unitaires ($n_i = 1$) and N_{nu} clusters non unitaires.

$$F(t,l) = (Q_1) \times (Q_2) = (\frac{N_u}{N}) \times (\frac{\max_{j \in 1..N}(n_j)}{\sum_{i=1}^{N} n_i}) \tag{3.1}$$

$$\begin{aligned} {t_1}^\theta &= argmin_t F(t,l) \\ l^\theta &= argmin_l F({t_1}^\theta, l) \end{aligned} \tag{3.2}$$

Minimiser F revient à minimiser Q_1 et Q_2.
Q_1 tend à maximiser le nombre de clusters non unitaires.
Q_2 assure un certain équilibre entre les tailles des clusters.

L'objectif de cette fonction est de s'assurer que les clusters résultants sont compacts et bien équilibrés au niveau taille, déterminant ainsi une structure hiérarchique optimale. Ainsi, l'algorithme proposé ne nécessite pas la connaissance préalable de seuils et permet la construction d'un VVH optimal d'une façon automatique.
Pour les niveaux intermédiaires, nous avons choisi d'augmenter le seuil du bas

1. chaque noeud de l'arbre possède n branches

vers le haut en doublant le pas entre deux niveaux consécutifs. Cette stratégie est motivée par le fait que les clusters qui sont en bas sont plus précis et donc plus important. Dans la figure 3.4, nous donnons une exemple de VVH de profondeur trois.

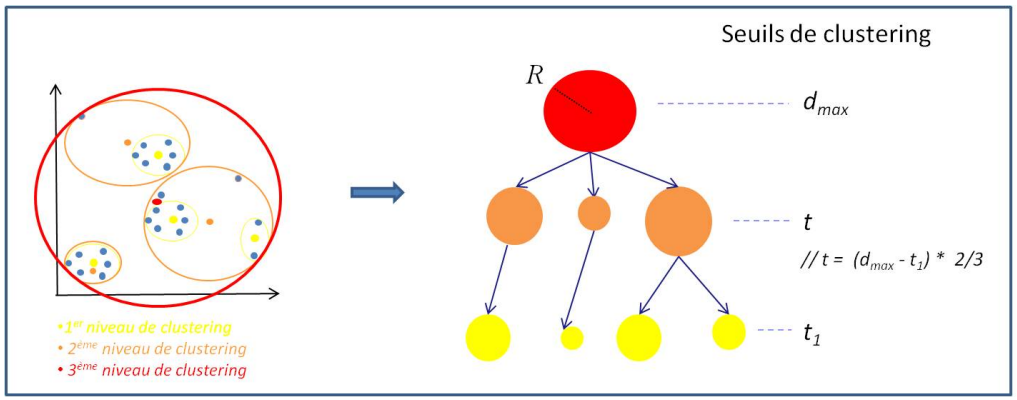


Figure 3.4. Un exemple de construction d'un Vocabulaire Visuel hiérarchique à trois niveaux

Une fois l'arbre construit, les clusters qui n'ont été fusionnés dans aucun niveau, sont supprimés. Ceci a pour objectif de condenser au maximum l'information sans pour autant perdre de sa pertinence.

3.2.3 Extraction d'une signature visuelle contenant des caractéristiques locales

L'extraction d'une signature visuelle se base sur la mise en correspondance des descripteurs SURF extraits d'une imagette de test avec le VVH construit en apprentissage sur une base d'images différente de celle du test. Ce processus est accéléré par l'exploitation de la représentation hiérarchique. En effet, une exploration partielle de l'arbre est généralement suffisante : lors de la mise en correspondance, il n'est pas nécessaire d'examiner les sous arbres dont le nœud père n'a pas été activé. Un nœud s'active si et seulement si la distance Euclidienne entre le descripteur du POI et le centroïde du cluster désigné est inférieure à son rayon R_i. Par conséquent, la mise en correspondance est réalisée en appliquant un simple algorithme itératif de parcours en profondeur. Le schéma récursif de mise en correspondance est représenté dans la figure 3.5.

Soit f_k un descripteur de POI SURF quelconque. Nous proposons d'attribuer une valeur d'activation $A_{i,k}$ du point k au cluster i par la formule suivante :

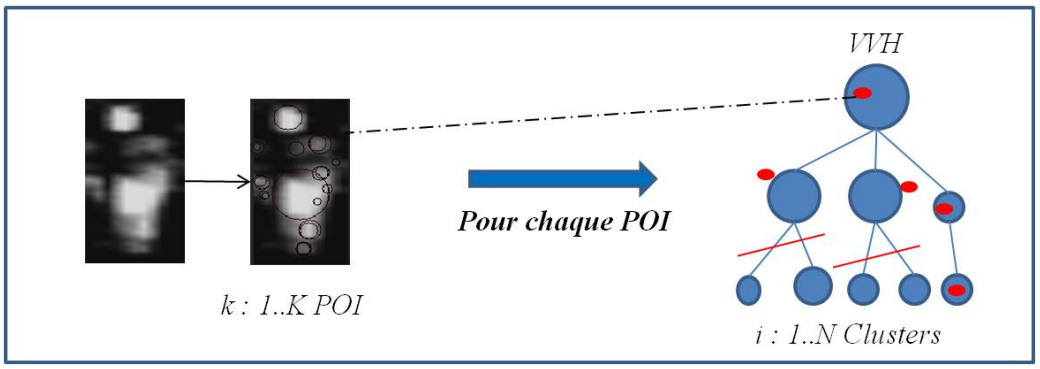


Figure 3.5. Schéma explicatif du processus de mise en correspondance entre un descripteur SURF et le VVH. Un gain approximatif de 50% au niveau du temps de calcul est réalisé vu que la moitié des nœuds du VVH n'ont pas été explorés

$$A_{i,k} = exp(-(\frac{c \times d(f_k, C_i)}{R_i})^2) \tag{3.3}$$

Avec c est un coefficient dont sa valeur a été fixée à 2 pour que la fonction $A_{i,k}$ génére des valeurs entre 0 et 1. Rappelons que R_i désigne le rayon du cluster i, C_i son centroïde et d représente la distance euclidienne dans l'espace des descripteurs. $A_{i,k}$ est une fonction décroissante de $d(f_k, C_i)$:

$$\begin{cases} d\,(f_k, C_i) > R_i \Rightarrow A_{i,k} \to 0 & (clusters\ non\ activés) \\ d(f_k, C_i) \leq R_i \Rightarrow A_{i,k} \in \texttt{]0,1]} & (clusters\ activés) \end{cases}$$

Cette formulation est également en accord avec le principe de l'exploration partielle de l'arbre. En effet, un nœud non activé ou non exploré se voit attribuer une valeur d'activation nulle. D'où tout l'intérêt de la structure hiérarchique qui permet d'accélérer considérablement le temps de mise en correspondance et de l'extraction des caractéristiques locales.

Afin de parvenir à l'invariance quant au nombre de POI extraits d'une imagette, nous proposons de normaliser toutes les valeurs d'activation (vote) avant leur accumulation dans l'histogramme. Plusieurs méthodes de normalisation ont été testées. Nous faisons une évaluation de ces méthodes dans la partie expérimentation. Ici, nous présentons la meilleure qui consiste à normaliser la valeur de chaque vote $A_{i,k}$ en fonction du niveau hiérarchique l en utilisant la formule suivante :

$$a_i = \frac{A_{i,k}}{\sum_{j=1}^{n_l} A_{j,k}} \tag{3.4}$$

Avec n_l est le nombre de clusters contenus dans le même niveau hiérarchique.

Ainsi, la valeur $\sum_{j=1}^{n_l} A_{j,k}$ désigne la somme totale de votes générés par la mise en correspondance du POI k avec l'ensemble de clusters j situés dans le même niveau hiérarchique que i.
Finalement, le vecteur caractéristique résultant et que nous proposons afin de constituer l'entrée du classifieur SVM est :

$$\mathbf{x} = (a_1, a_2, \ldots, \ldots, a_N)$$

Avec N est le nombre total de clusters contenus dans le VVH.
Nous récapitulons le principe général de l'extraction de la signature visuelle d'apparence locale dans la figure 3.6.

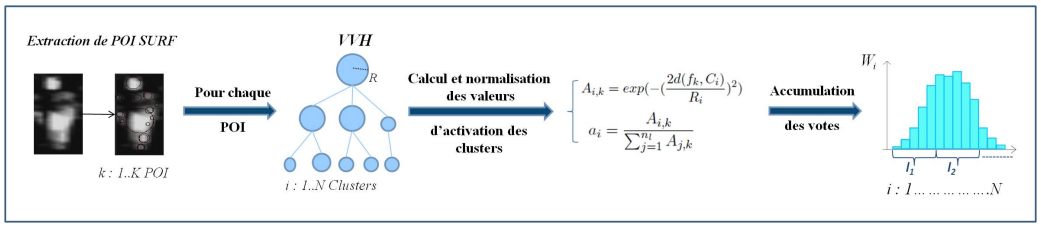


Figure 3.6. Extraction d'une signature visuelle qui caractérise l'apparence locale d'un piéton en utilisant le VVH

La figure 3.6 résume les étapes essentielles de l'extraction d'une signature visuelle en allant de l'étape d'extraction de POI, jusqu'à la construction d'un histogramme qui caractérise l'apparence locale d'un objet.

3.3 Catégorisation par combinaison du VVH avec des méthodes à noyaux

L'utilisation de VV pour représenter les apparences locales s'est montrée très efficace dans la résolution des problèmes d'occultations partielles et de changements de point de vue [LLS04, Lei08, LBH08]. Cette technique est considérée comme référence pour gérer les grandes variations de formes et d'apparences intra-classe. Néanmoins, la méthode présente le désavantage d'être basée sur une représentation très flexible, qui est plus susceptible d'être sujet à des faux positifs. Cela prouve bien la nécessité d'intégrer des techniques de discrimination pour améliorer le système de catégorisation automatique. Ces dernières années, il y'a eu un intérêt croissant dans le développement d'algorithmes qui combinent des représentations locales d'image avec des systèmes à base de méthodes à noyaux comme les SVMs [WCG03, FLCS05, LBH08]. Sur ce problème, les noyaux permettent de définir des

mesures de similarité entre des ensembles de descripteurs.
Nous distinguons deux manières d'utiliser le VV pour comparer des listes de descripteurs locaux, les distances inter-histogramme et les noyaux par mise en correspondance (LMK : Local Matchnig Kernel).

3.3.1 Noyaux pour histogrammes

L'histogramme est un mode de représentation simple et particulièrement répandu. Cependant, la comparaison d'histogrammes pose un problème dans le cas d'une forte proportion de valeurs nulles. Pour remédier à ce problème, nombreuses mesures de similarité permettant la comparaison d'histogrammes ont été proposées. Nous ne citerons ici que les plus utilisées qui sont basées sur une distance d'intersection d'histogrammes ou de χ^2.
Le noyau d'intersection d'histogrammes se base sur le calcul de la distance d_{IH} entre deux histogrammes [], définie telle que :

$$d_{IH}(H1, H2) = \sum_{i=1}^{N} min^{\beta}(h_i, h'_i), \beta \geq 0 \tag{3.5}$$

Avec h_i et h'_i sont respectivement les valeurs du bin i dans les histogrammes $H1$ et $H2$.
Le d_{χ^2} est une mesure statistique qui est généralement utilisée pour effectuer des tests d'hypothèses. Cette mesure a été adaptée à la comparaison d'histogrammes :

$$d_{\chi^2}(H1, H2) = \sum_{i=1}^{N} \frac{(h_i - h'_i)^2}{(h_i + h'_i)} \tag{3.6}$$

La distance d_{χ^2} est généralement utilisée comme fonction de distance pour un noyau RBF ($K_{d_{\chi^2}} = \exp^{(-\frac{d_{\chi^2}}{\sigma})}$).

3.3.2 Noyaux par mise en correspondance (LMK)

Le noyau LMK a été proposé par [WCG03]. Le principe est de comparer les imagettes par la mesure d'une similarité sur tous les couples possibles de descripteurs issus de chacune des deux imagettes. Cette similarité entre descripteurs est binaire. Elle vaut 1 s'il existe un élément unique du dictionnaire décrivant au mieux chacun des deux descripteurs à comparer.
Soit X et Y deux ensembles de descripteurs avec $X = \{x_i\}_{i=1}^{n1}$ et $Y = \{y_j\}_{j=1}^{n2}$. Le noyau LMK est définit par :

$$K_L(X,Y) = \frac{1}{2}[K(X,Y) + K(Y,X)] \tag{3.7}$$

Le noyau $K(X,Y)$ permet de fournir la moyenne de la meilleure mise en correspondance des scores correspondants aux éléments X et Y :

$$K(X,Y) = \frac{1}{n1}\sum_{i=1}^{n1} \max_{j=1..n2} K_l(x_i, y_j) \tag{3.8}$$

Avec $n1$ et $n2$ sont le nombre des POI détectés des deux imagettes, K_l (Local Kernel) est un noyau défini dans l'équation 3.9.
Inspiré par cette formulation, [FLCS05] a adapté la formulation du noyau au Vocabulaire Visuel en exprimant le produit scalaire $< \vec{x}, \vec{y} >$ en terme d'un problème de mise en correspondance avec le VV.
Soit $X = ((\vec{x_1}, \lambda_1), ..., (x_{n1}, \vec{\lambda_{n1}}))$ un ensemble de descripteurs locaux où x_i représente le descripteur d'apparence et λ_i l'emplacement relatif du POI. Soit $A = (\vec{A_1}, ..., \vec{A_{n1}})$ le vecteur codant les activations des descripteurs pour l'ensemble des N clusters du VV, avec $\vec{A_i} = (a_1, ..., a_N)$. Pour chaque paire de vecteur descripteurs $(\vec{x}, \lambda_x)$ et $(\vec{y}, \lambda_y)$, la mise en correspondance est évaluée par la mesure de similarité locale :

$$K_l((\vec{x}, \lambda_x), (\vec{y}, \lambda_y)) = K_a(\vec{x}, \vec{y}) \times K_p(\lambda_x, \lambda_y) \tag{3.9}$$

Où K_a, K_p représentent une mesure de similarité associée à l'apparence et respectivement l'emplacement.

$$\begin{aligned} K_a(\vec{x}, \vec{y}) &= \exp(-\Upsilon(1 - < \vec{x}, \vec{y} >)) \\ &\approx exp(-\Upsilon(1 - \textstyle\sum_i \sum_j a_i < \vec{C_i}, \vec{C_j} > b_j)) \end{aligned} \tag{3.10}$$

Notons que b_j, pareillement à a_i, est un binaire codant l'activation du cluster j. Υ est un coefficient de relâchement.

$$K_p(\lambda_x, \lambda_y) = exp(-\frac{(\lambda_x - \lambda_y)^2}{2\rho^2}) \tag{3.11}$$

Soit $\phi \in \pi_1^{n1}$ et $\psi \in \pi_1^{n2}$ désignant les permutations possibles de descripteurs, le LMK correspondant est alors défini comme suit :

$$K(X,Y) = \frac{1}{k}\max_{\phi,\psi}\sum_{j=1}^{k} K_l((\vec{x}_{\phi(j)}, \lambda_{x_\phi(j)}), (\vec{y}_{\psi(j)}, \lambda_{y,\psi(j)})) \tag{3.12}$$

D'après les équations 3.10 et 3.12, la complexité d'un noyau défini entre deux objets est de $\mathcal{O}(n1.n2.N^2)$. Cette représentation, bien que pertinente, nécessite le calcul de toutes les correspondances entre descripteurs ; ce qui implique une grande complexité combinatoire. De plus, la partie multiplicative du noyau possède deux paramètres de largeur de bande à optimiser. Les paramètres peuvent être trouvés au moyen d'une validation, mais le grand nombre d'hyperparamètres implique une grille fine de validation et donc un processus fastidieux de recherche pouvant amener au sur-apprentissage. Dans tous les cas, il est important de comparer les performances des différentes formes de noyaux sur des bases communes d'OR.

3.3.3 Expérimentations et évaluations

Les systèmes dotés de la technologie infrarouge-lointain embarquée sur un véhicule, sont principalement destinés à la détection de piétons. À ce sujet, on trouve plusieurs travaux dans la littérature [SRBB06, BBF+04, BBG+07, JA09]. En revanche, très peu ont abordé la problématique de détection de véhicules [AAB+02]. Cela se justifie par le fait que les véhicules, notamment stationnés, ne présentent pas un contraste suffisant pour être visualisés dans les images IR.
Pour ces raisons, nous avons choisi d'optimiser et d'évaluer, tout d'abord, les paramètres et les performances du système face au problématique de la reconnaissance de piétons. Le problème de généralisation et les performances de reconnaissance génériques des OR sont représentés dans la section 3.4.
Les expérimentations sont faites sur une base annotée de piétons en IR extraites du système de Tetravision (voir section 1.4.2). Les courbes ROC (Receiver Operating Characteristic) ont été utilisées pour évaluer les paramètres des modèles proposés. En effet, l'air sous ces courbes (AUC) constitue un outil majeur d'évaluation des résultats. Ces courbes expriment le pourcentage de faux positifs en fonction du pourcentage des vrais positifs. Les courbes ROC permettent d'évaluer un problème de classification mono-classe. Ainsi, nous avons évalué les performances de classification multiclasses à partir d'autres mesures très connues :

- le taux de reconnaissance, TR = $\frac{TP+TN}{TP+FP+FN+TN}$,
- la précision, P = $\frac{TP}{TP+FP}$,
- le rappel, R = $\frac{TP}{TP+FN}$,
- la F-mesure, (F-score)= $\frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2.TP}{2.TP+FN+FP}$.

Avec TN, TP, FN, FP désignent respectivement les nombre de vrai négatifs, vrai positifs, faux négatifs et faux positifs.
Il est important de noter que tous les résultats que nous allons présenter dans cette

section serviront à valider les différents paramètres du système de reconnaissance proposé. Ainsi, nous présenterons les résultats d'expérimentations obtenus sur la base de validation *Tetra5* contenant 2111 objets (1089 piétons,1003 non piétons).

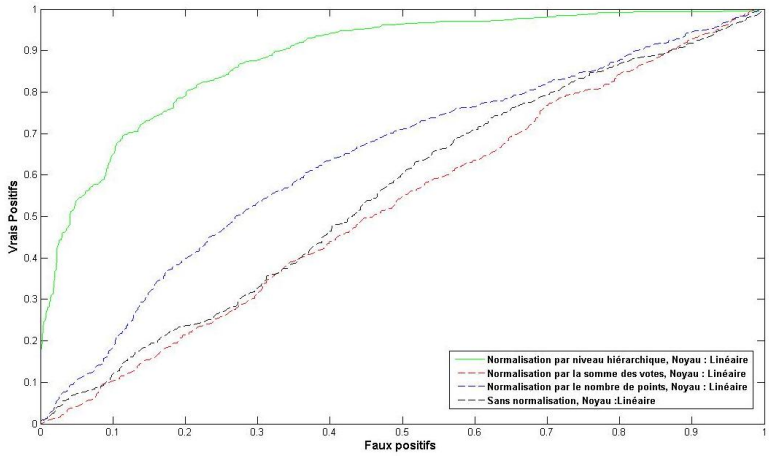
3.3.3.1 Evaluation de la méthode de normalisation des votes

La caractérisation des images par des descripteurs invariants à l'échelle ne permet que de résoudre une partie du problème de reconnaissance. En effet, le nombre de POI SURF extraits d'une image varie en fonction de sa résolution. La méthode proposée pour normaliser les votes générés lors du processus de mise en correspondance constitue une alternative à la tâche de redimensionnement de fenêtres qui demande un temps de calcul non négligeable.
Mise à part la méthode de normalisation (Eq 3.4) que nous avons proposée dans la section 3.2.3 (1), les autres méthodes envisagées sont les suivantes :
- (2) Normalisation par la somme des activations de clusters contenus dans l'arbre (inclure tous les niveaux hiérarchiques).
- (3) Normalisation de tous les bins de l'histogramme après génération, par le nombre total de POI extraits de l'objet en question.
Dans la figure 3.7, nous évaluons chacune de ses méthodes en comparant les résultats de classification obtenus avec un SVM à noyau linéaire et respectivement RBF.

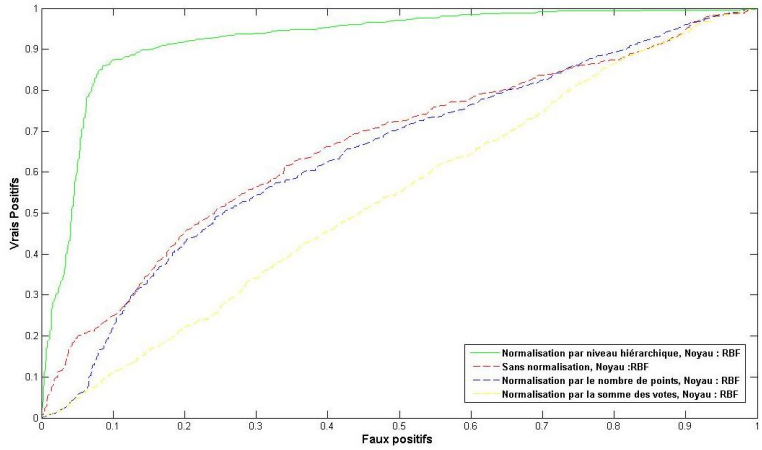
Les courbes obtenues montrent que la méthode de normalisation a une grande importance. Il est clair que l'Aire sous la courbe verte, correspondante à la normalisation par niveau hiérarchique, est nettement au dessus des autres courbes, à savoir la méthode (2) et (3). Cela nous amène à dire que la normalisation par niveau hiérarchique est plus adaptée à la structure hiérarchique du VV. En outre, il est clair que les résultats obtenus avec un noyau linéaire sont aussi bons que ceux obtenus avec un noyau RBF. Dans la section suivante, nous évaluons différentes fonctions noyaux pour la classification.

3.3.3.2 Choix des fonctions noyaux

Dans cette section, nous évaluons les différentes fonctions noyaux afin de retenir la plus adaptée au modèle de représentation proposé. Le noyau linéaire ne nécessite aucun paramétrage. Les autres fonctions, à part le noyau LMK, n'utilisent qu'un seul paramètre. Pour le noyau d'intersection d'histogrammes, il s'agit du facteur de puissance β. Pour les noyaux RBF et RBF-χ^2, il s'agit du paramètre de largeur de bande. Quant au noyau LMK, il inclut d'autres paramètres qui sont les valeurs de



(a)



(b)

Figure 3.7. Résultats de classification obtenus avec les différentes méthodes de normalisation de vote lors du processus de mise en correspondance. Les courbes ROC ont été obtenues en utilisant un noyau linéaire dans (a) et un noyau RBF dans (b).

Υ et de ρ. La valeur du paramètre de pénalisation C et les paramètres spécifiques du noyau ont été validés en cherchant les valeurs optimales sur une grille de deux dimensions (voir section 2.2.2.1).

Les courbes ROC présentées dans la figure 3.8 montrent que les meilleures performances sont obtenues avec un noyau RBF. L'utilisation de cette fonction noyau permet d'avoir 92.5% comme valeur d'AUC. Les autres valeurs d'AUC fournies, par ordre décroissant, sont : 92.1% (intersection d'histogrammes), 91.3% (RBF$-\chi^2$), 88.7% (LMK), 84.5% (linéaire).

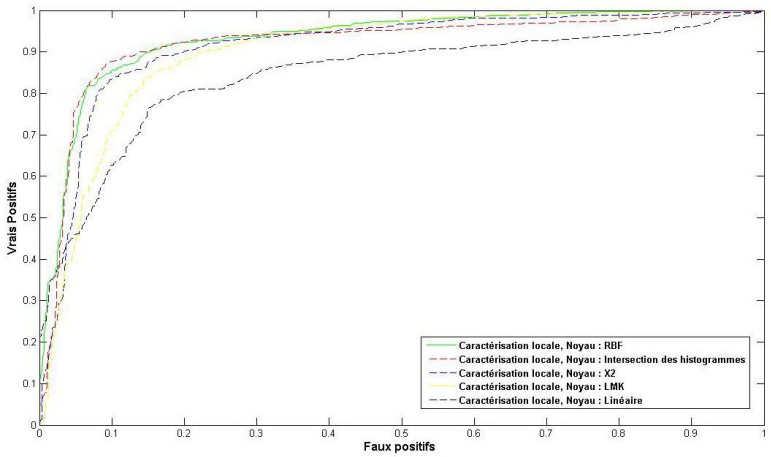


Figure 3.8. Présentation des courbes ROC pour l'évaluation de fonctions noyaux pour les caractéristiques locales

Le gain de performances obtenu avec le noyau RBF par rapport au noyau linéaire est expliqué par la capacité du premier à approximer des séparateurs très complexes. Le noyau linéaire obtient des résultats corrects, mais dépassés par toutes les autres fonctions présentées. Les noyaux d'intersection d'histogrammes et de RBF$-\chi^2$ sont généralement utilisés pour des mesures de similarité entre histogrammes d'accumulation de votes binaires. Tandis que, la méthode que nous avons proposée pour l'extraction de caractéristiques, bien qu'elle utilise des histogrammes, elle accumule des votes non binaires, qui sont normalisés par niveau hiérarchique.
De même pour les caractéristiques globales, nous comparons dans la figure 3.9 les résultats obtenus en fonction des noyaux utilisés. Vu que les caractéristiques globales ne présentent aucune particularité, telle qu'une structure hiérarchique ou l'utilisation d'histogrammes, nous ne présentons que les courbes spécifiques à l'utilisation des noyaux linéaire et RBF.

Les courbes présentées dans la figure 3.9 montrent de bons résultats obtenus en utilisant les caractéristiques globales. Comme précédemment constaté, les meilleures performances sont obtenues avec un noyau RBF. Ces résultats nous amèneront ainsi à discuter l'apport de la fusion des caractéristiques locales et globales en utilisant le noyau RBF (voir figure 3.10).

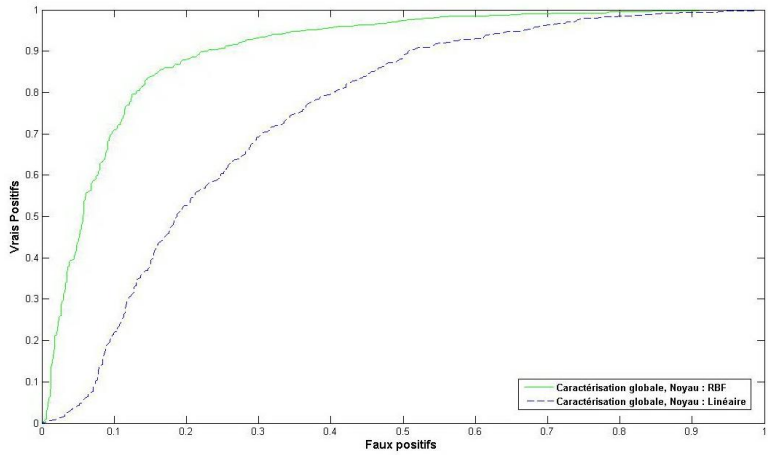


Figure 3.9. Présentation des courbes ROC pour l'évaluation des fonctions noyaux pour les caractéristiques globales

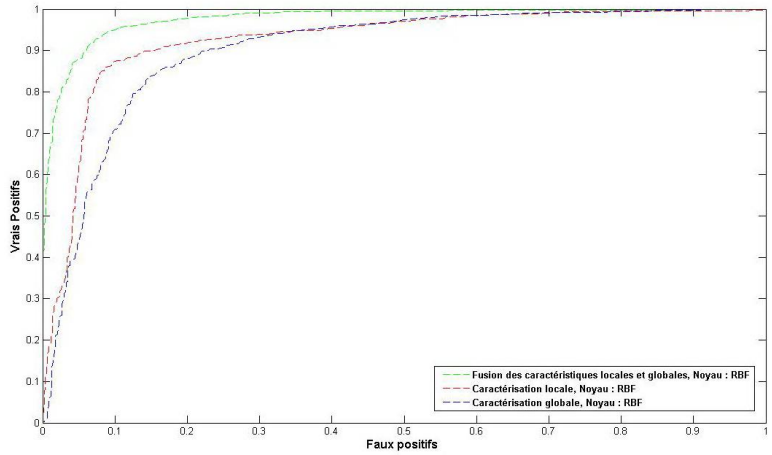


Figure 3.10. Présentation des courbes ROC pour l'évaluation de la fusion de caractéristiques locales et globales

3.3.3.3 Evaluation de la fusion des caractéristiques

Dans la figure 3.10, nous présentons les courbes ROC correspondantes à une représentation locale, une représentation globale et respectivement une fusion de représentation locale et globale. Le processus de fusion de caractéristiques consiste à concaténer les caractéristiques locales et globales dans un seul vecteur caractéristique. Ce vecteur a été normalisé avant d'être utilisé en entrée de SVM à noyau RBF. La normalisation consiste en une transformation gaussienne afin que les attributs suivent une distribution normale centrée réduite (de moyenne 0 et variance 1).

La comparaison des Aires sous les courbes ROC montre que les caractéristiques locales (AUC=92.5%) sont légèrement plus pertinentes que les caractéristiques globales (AUC=90.1%). Cela pourrait être dû par le fait que le nombre de caractéristiques locales utilisées est largement plus grand (159 caractéristiques) que le nombre de caractéristiques globales (30 caractéristiques). Ce qui est plus important est que la fusion de caractéristiques a pu améliorer les résultats de plus que 5% (AUC=97.6%). D'où tout l'intérêt du processus de fusion qui permet d'enrichir et de fiabiliser les données.

3.3.3.4 Evaluation des descripteurs SURF Vs SIFT

Les principales briques du système de reconnaissance proposé sont basées sur l'extraction et la mise en correspondance de descripteurs. Ainsi, le choix du descripteur est d'une importance cruciale. Tous les résultats présentés jusqu'à présent ont été obtenus avec un descripteur de type SURF-64. Bien évidemment, des considérations de temps de mise en correspondance sont à l'origine de ce choix. Dans la figure 3.11, nous comparons les résultats obtenus en utilisant les descripteurs SURF-64, SURF-128 et SIFT.

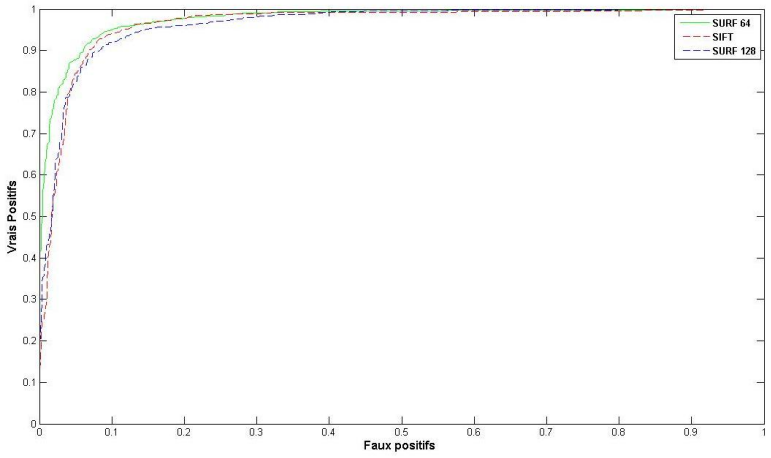


Figure 3.11. Présentation des courbes ROC pour l'évaluation des descripteurs SIFT,SURF-64 et SURF-128

Les courbes présentées dans la figure 3.11 montrent que les descripteurs SURF sont plus adaptés que le SIFT dans le contexte de reconnaissance de piétons à partir des images IR. Cela confirme les analyses que nous avons faites dans la section 3.1.1.
Bien que la description SURF-128 soit plus riche, les résultats obtenus en utilisant

SURF-64 sont meilleurs. Deux explications possibles de cette divergence sont suggérées. D'une part, les VVHs établis en apprentissage sont différents et n'ont pas la même taille. Il est possible que l'augmentation de la taille des descripteurs conduise à générer des clusters qui ne sont pas pertinents pour la classification. D'autre part, nous considérons que la différence entre SURF-128 et SURF-64 (voir section 2.1.3.2) ne produit pas d'information particulièrement pertinente. Par conséquent, nous avons retenu le descripteur SURF-64.

3.3.3.5 Evaluation de la structure du VVH

Les performances de reconnaissance sont largement dépendantes de la profondeur du VVH. Dans la section 3.2, nous avons proposé une fonction qui permet d'évaluer la pertinence de la structure du VV. Bien que, durant la phase de validation, il a été constaté qu'un VVH de profondeur 5 permet de minimiser la fonction d'évaluation F. Nous présentons dans la figure 3.12 l'évolution des performances de reconnaissance en fonction de la profondeur du VVH.

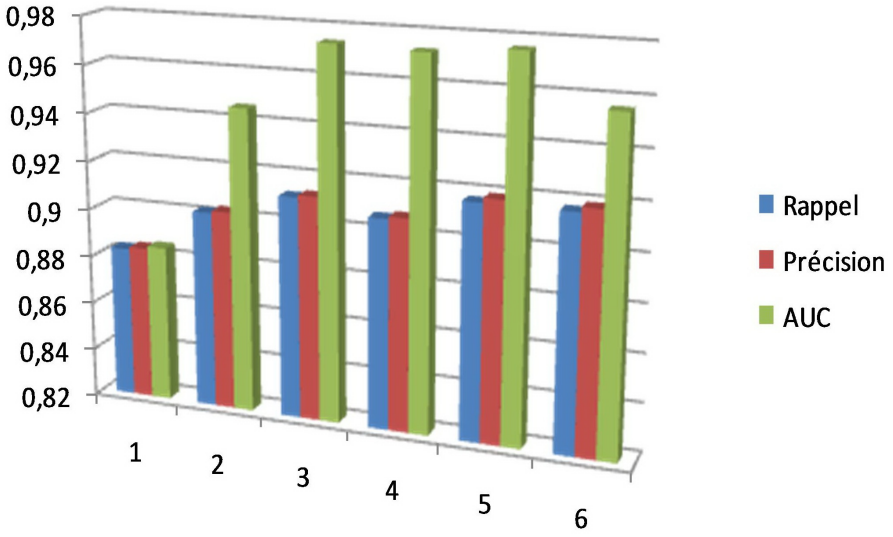


Figure 3.12. Evolution des performances de reconnaissance en fonction de la profondeur du VVH

Dans la figure 3.12, nous observons que les résultats obtenus diffèrent significativement en fonction de la profondeur du VVH. Nous voyons également que les meilleurs résultats sont obtenus avec l'utilisation de 5 niveaux hiérarchiques. Cela confirme la pertinence de la fonction d'évaluation proposée. En ce qui concerne l'évolution des résultats, la courbe montre que des améliorations significatives ont

été apportées grâce à la structure hiérarchique du VV. En effet, jusqu'à la profondeur 5, les résultats continuent à s'améliorer. Cela prouve que l'augmentation de la profondeur de l'arbre jusqu'à un certain niveau permet d'améliorer la pertinence des caractéristiques extraites pour la classification SVM. Ainsi, le fait d'inclure plusieurs valeurs de seuils dans la structure hiérarchique permet d'augmenter le nombre de clusters non unitaires dans le VVH. En revanche, à partir d'une certaine profondeur, les résultats commencent à se dégrader vu que des clusters non significatifs commencent à apparaître dans le VVH. Cette dégradation peut être liée également au comportement des SVM pour des vecteurs d'entrée en haute dimension.

3.3.4 Bilan

Grâce aux expérimentations effectuées toute au long de cette partie, nous avons parvenu à fixer les paramètres les plus adaptés à la tâche de reconnaissance. Tous les aspects liés aux méthodes de construction du VVH, d'extraction des caractéristiques, de normalisation et du choix du descripteur et du noyau ont été étudiés. Sur nos bases d'images, le système présente des résultats optimaux pour la combinaison du VVH, appris sur des descripteurs SURF-64, avec un SVM utilisant un noyau RBF. Les résultats obtenus mettent notamment en valeur le processus de la fusion des caractéristiques locales et globales qui a permis d'améliorer significativement les résultats de la reconnaissance de piétons. Dans la section suivante, nous présentons les résultats d'expérimentation obtenus sur une base d'images multiclasses incluant des exemples de véhicules et de fonds d'images.

3.4 Evaluation des performances de reconnaissance multiclasses des obstacles routiers

Dans cette section, nous présentons des résultats expérimentaux obtenus sur la base multiclasses *Tetra6* (voir section 1.4.2). L'ensemble des données comprend un total de 986 objets répartis entre des piétons, véhicules et fond d'images. Il est important de mentionner que la diversité des exemples est très importante : la variance des échelles est de 3,8 et presque 17% des piétons et des véhicules sont partiellement occultés. Notons que le taux d'occultation (rapport intersection/union de chevauchement) considéré est de 60%. La figure 3.13 présente un extrait de quelques objets annotés d'une image IR prise de cette base.

Le principe général du système de reconnaissance est décrit dans la figure 3.14.

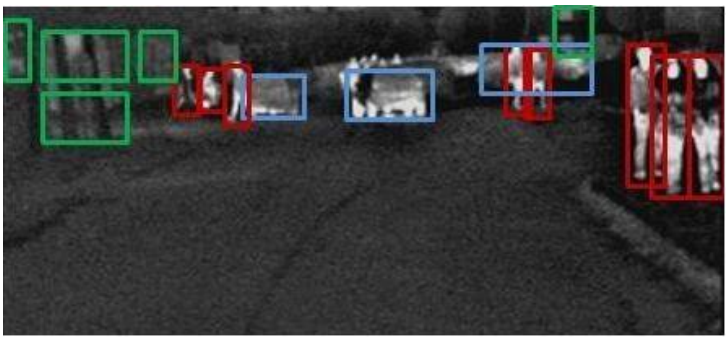


Figure 3.13. Extrait de quelques objets annotés d'une image IR. Les piétons, les véhicules et le fond sont encadrés par des fenêtres englobantes de couleurs rouge, bleu et vert.

Toutefois, il est important de souligner les extensions faites afin que le système proposé puisse traiter un problème de reconnaissance multiclasses. En ce qui concerne la représentation, un VVH a été construit pour chaque classe d'objets (Piéton, Véhicule, Fond d'image). Ainsi, le processus d'extraction de caractéristiques locales procède par la mise en correspondance de 3 VVH. De ce fait, chaque POI extrait d'une imagette doit parcourir le VVH de Piéton, de Véhicule et de Fond d'image. Ensuite les trois signatures sont concaténées constituant un seul vecteur caractéristiques. Concernant la classification, un SVM multiclasses, utilisant l'approche un-contre-un, a été utilisé. Ces points sont détaillés dans la section suivante.

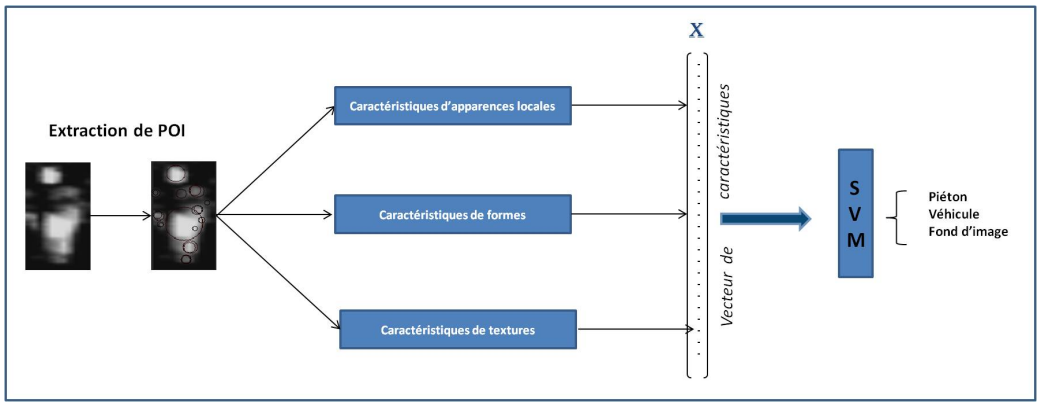


Figure 3.14. Description générale du système de reconnaissance

3.4.1 Apprentissage du modèle

Tout d'abord, nous avons extrait de la base d'apprentissage un sous-ensemble afin de construire un VVH de piéton, un VVH de véhicule et un VVH de fond d'image. Cet ensemble contient seulement des exemples d'objets non occultés,

mais qui présentent une grande variabilité d'apparences et de formes. Ensuite, un classifieur SVM a été appris sur l'ensemble total d'apprentissage qui contient 671 imagettes incluant des situations d'occultations. La figure 3.15 montre quelques exemples d'images utilisées pour l'apprentissage de SVM. Il est important de mentionner que les images présentées dans la figure sont redimensionnées juste pour des raisons de clarté de présentation. Vu que l'ensemble des caractéristiques proposées sont invariantes à l'échelle, notre système ne nécessite pas une étape préalable de normalisation des résolutions.

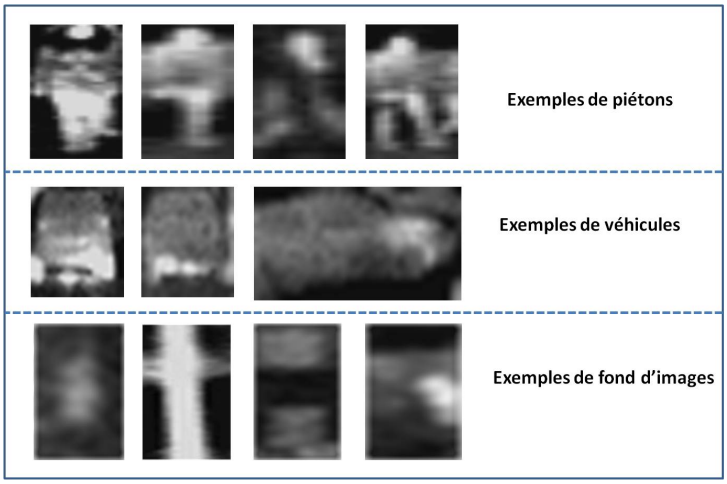


Figure 3.15. Exemples d'images d'apprentissage pour les classes piéton, véhicule et fond d'image

Pour l'apprentissage de SVM, nous avons utilisé une technique de validation croisée de 10 itérations afin de faire le choix du noyau et d'optimiser ses hyperparamètres. Les résultats optimaux ont été trouvés en utilisant un noyau RBF avec $C = 10^3$ et $\gamma = 0.001$. Ces résultats sont donnés dans le tableau 3.1.

Table 3.1. Résultats de classification obtenus sur la base de validation

Caractéristiques	Locales	Globales	Fusion
Nombre de caractéristiques	286	30	316
Taux de reconnaissance (TR) (%)	63.94	87.41	91.21
F-score (%)	63	87.5	91.2
AUC (%)	74.1	91.8	97.2

Le tableau 3.1 montre qu'une amélioration de 7% de la valeur AUC a été réalisée par la combinaison des caractéristiques locales et globales. Notons bien qu'à la différence des résultats obtenus en mono-classe (voir figure 3.10), les caractéristiques globales sont plus pertinentes pour la discrimination multiclasses. Cela

montre que la caractérisation des apparences locales n'est significative que pour la catégorie piéton. En effet, ce n'est que dans les images de piétons où des régions spécifiques apparaissent contrastés (Par exemple, présence de régions claires autour des têtes et des jambes de piétons).

3.4.2 Classification

Dans cette section, nous donnons des résultats obtenus sur une base de test contenant 297 objets représentant des piétons, des véhicules et des fonds d'images. Les résultats obtenus seront confrontés avec des résultats de classification de caractéristiques références comme les ondelettes de Haar et de Gabor. En outre, nous étudions dans cette section le problème de la réduction du temps de la classification. Ayant validé le choix du noyau dans les sections précédentes, nous examinons ici l'impact du processus de sélection d'attributs. Bien que les résultats de fusion de caractéristiques précédemment présentés sont très satisfaisants, la concaténation des vecteurs caractéristiques locales et globales conduit à un espace de grande dimension. Cela amène à un temps de classification très élevé. Pour remédier à ce problème, nous proposons de réduire le nombre de caractéristiques en appliquant un algorithme de sélection d'attributs. Ainsi, nous proposons d'intégrer une étape de sélection des attributs les plus pertinents. Dans la section, 2.3.1, nous avons présenté plusieurs techniques et mesures d'évaluations. Après plusieurs tests de validations, nous avons choisi d'appliquer l'algorithme proposé dans [Hal98]. Cet algorithme, très rapide, permet de sélectionner les attributs les plus pertinents en évaluant le taux de corrélation entre des sous ensembles d'attributs. Dans le tableau 3.2, nous menons une comparaison entre les résultats de classification obtenus avec et sans processus de sélection de caractéristiques.

Table 3.2. Résultats de classification obtenus sur la base de test

Classe	Caractéristiques	Piéton	Véhicule	Fond d'images	Moyenne
F-score (%)	Sans sélection	95	90.1	89.7	**91.6**
	Avec sélection	94.7	86.5	88.8	**90**
AUC (%)	Sans sélection	96.9	91.2	95.7	**94.6**
	Avec sélection	98	88.2	95.4	**93.87**

Les résultats présentés dans le tableau 3.2, montrent que la reconnaissance de la catégorie de piétons est plus facile que celle de véhicules ou de fond d'image. Cette constatation s'explique par le fait qu'un contraste thermique significatif ca-

ractérise la présence de piétons dans les images IR. En revanche, cette propriété n'est pas forcément vérifiée pour les autres objets, notamment les véhicules qui ne sont pas en mouvement. Toutefois, le tableau montre que notre système de reconnaissance fournit des résultats satisfaisants, présentant une moyenne de 95% d'AUC. Il est intéressant de constater aussi que l'intégration du processus de sélection de caractéristiques (réduction du nombre de 316 à 24) avant classification permet d'atteindre 94% d'AUC. Ainsi, les résultats obtenus sans (SC) ou avec (AS) sélection de caractéristiques sont très proches. Par conséquent, nous pouvons considérer que la sélection de caractéristiques a permis d'éliminer les attributs redondants tout en conservant les performances de classification. Nous donnons dans le tableau 3.3 des informations sur les temps de classification. Ces informations montrent que le processus de SC permet de réduire jusqu'à 7 fois le temps de calcul nécessaire pour la classification.

Table 3.3. Tableau comparatif des performances de classification

	Notre méthode		Ondelettes de Haar	Ondelettes de Gabor	Haar + Gabor
	SS	AS			
Nombre de caractéristiques	316	**24**	64	32	96
Taux de reconnaissance (%)	**91.51**	89.8	71.76	84.69	89.11
F-score (%)	**91.6**	90	72.2	84.5	89
AUC (%) (%)	**94.6**	93.87	83.6	90.2	93.3
Temps de classification (ms)	4.84	**0.7**	4.5	1.7	1.9

Dans le tableau 3.3, nous présentons les performances globales du système de reconnaissance proposé. Tous les résultats exposés sont obtenus en utilisant un SVM avec des hyperparamètres optimisés. Il est important de mentionner aussi que contrairement à notre méthode, les ondelettes de Haar et de Gabor ont été extraites après normalisation de la résolution des imagettes.

Dans le chapitre 2, nous avons présenté un état de l'art sur les méthodes de représentation et de caractérisation des obstacles routiers. Les caractéristiques extraites à partir des ondelettes de Gabor et de Haar ont été bien détaillées vu qu'elles sont considérées comme des méthodes de références. La première permet de caractériser globalement l'image, tandis que, la deuxième assure une caractérisation par région. Nous rappelons que le modèle de représentation proposé dans notre sys-

tème permet de caractériser non seulement localement, mais aussi globalement les images. Pour pouvoir juger sa pertinence, nous comparons les résultats globaux obtenus avec les deux méthodes références.
La comparaison de nos résultats avec ceux obtenus en utilisant les caractéristiques de Haar et Gabor, montre que notre méthode de représentation est plus discriminative. De nombreux auteurs ont prouvé que les ondelettes de Haar et Gabor offrent des informations complémentaires et que leur fusion permet d'améliorer les performances de reconnaissance [SBM05, ARB08]. Les résultats présentés dans le tableau montrent que notre système, bien qu'il soit basé sur une représentation très compacte (24 caractéristiques), donne des résultats meilleurs qu'une représentation fondée sur la fusion de caractéristiques de Haar et de Gabor. Cela prouve en partie que toute méthode, considérant l'intégralité des pixels de l'image afin d'extraire une signature, ne permet pas de résoudre les problèmes liés à l'occultation et la présence majoritaire de pixels non-objet.
Pour conclure, les résultats présentés dans cette section prouvent que notre méthodologie SURF/SVM est capable de résoudre des taches complexes de reconnaissance. D'une part, la représentation d'objet par un ensemble de POI SURF permet de répondre aux problèmes posés par les larges variations de formes et d'apparences, et des occultations partielles de l'objet. D'autre part, l'utilisation de SVM permet de prendre en compte et de discriminer la configuration globale des caractéristiques.

3.4.3 Bilan

Dans cette section, nous avons évalué les performances multiclasses de notre système de reconnaissance. Le système s'appuie d'un côté sur la mise au point d'un modèle de représentation basé sur des caractéristiques locales et globales et, de l'autre, sur la classification par SVM à noyau RBF. Les résultats obtenus montrent que le système proposé devance ceux basés sur d'autres modèles de représentation (ensemble de POI SIFT, Ondelettes de Haar et de Gabor) ou ceux utilisant d'autres fonctions noyaux (LMK) pour la classification.
Un autre point particulièrement important qui a bien été mis en valeur est l'apport du processus de fusion de caractéristiques locales et globales. Les caractéristiques locales et globales proposées ont fait preuve de complémentarité. Il semble donc intéressant de poursuivre l'étude de l'apport de fusion en utilisant d'autres techniques plus avancées. Ainsi, nous proposons d'étudier l'implantation d'un module de fusion multimodale pour tirer profit non seulement de la complémentarité des

caractéristiques locales et globales extraites depuis des images IR, mais aussi des modalités VIS et IR. Nous aborderons cela dans le chapitre suivant.

3.5 Conclusion

Dans ce chapitre, nous avons présenté notre contribution concernant la représentation et la catégorisation des obstacles routiers. Dans un premier temps, nous avons proposé un Vocabulaire Visuel Hiérarchique construit à partir d'un ensemble de points d'intérêt. Ce modèle permet de caractériser les apparences locales d'une façon flexible. Néanmoins, cette flexibilité, bien qu'elle soit bien adaptée aux larges variations des apparences locales, a tendance à faire générer des faux positifs. Ainsi, nous avons proposé dans un deuxième temps de fusionner les caractéristiques locales avec des caractéristiques globales. Enfin, nous avons proposé une méthode permettant de combiner notre modèle de représentation avec la technique d'apprentissage supervisée SVM.
Différentes observations expérimentales ont été menées afin d'étudier le choix des composants du système de reconnaissance proposé. En outre, nous avons confronté les résultats obtenus à des méthodes de représentation et de catégorisation de référence dans la littérature. Les résultats expérimentaux obtenus montrent l'intérêt de notre système de reconnaissance qui présente de bons résultats en IR. Dans le chapitre suivant, nous étudions l'intégration d'un module de fusion multimodale afin d'améliorer les performances globales du système de reconnaissance.

Chapitre 4

Fusion multimodale pour la reconnaissance des obstacles routiers

Introduction

Dans les chapitres précédents, nous avons mis en évidence la notion de fusion d'informations pour améliorer l'analyse et l'interprétation des données. Disposant de données hétérogènes, issues de capteurs différents, la fusion pourrait également constituer l'étape ultime et décisionnelle d'un système de reconnaissance.
Dans ce chapitre, nous commençons par justifier le cadre de fusion retenu avant de présenter ses outils et ses mécanismes de base. Ensuite, nous présentons les méthodes proposées et les stratégies mises en jeu afin d'implémenter un système fiable et rapide de reconnaissance d'OR. À la fin du chapitre, nous donnons des résultats expérimentaux obtenus sur des images visibles et infrarouges issues de scènes routières.

4.1 Contexte d'application

La fusion multimodale des images a été explorée et développée pour plusieurs applications. L'aide à la conduite pour la rendre plus sécuritaire est l'une de ces applications d'importance. La fusion d'images visibles et infrarouges peut contribuer à améliorer grandement la performance d'un système de reconnaissance des OR. En prenant l'avantage de la complémentarité des deux modalités, la fusion peut assurer l'amélioration de l'analyse et de l'interprétation des données fournies par les capteurs.

Les résultats de reconnaissance obtenus dans le chapitre précédent nous conduisent à souligner plusieurs points importants. Tout d'abord, chaque modalité en soi ne peut pas être utilisée de manière fiable pour effectuer la reconnaissance. Ainsi, on peut s'attendre à intégrer un module de fusion multimodale pour que la reconnaissance soit plus performante.
Les sources d'informations à fusionner sont liées à deux catégories de caractéristiques : locale (L) et globale (G) qui sont issues, à leur tour, de deux modalités différentes (VIS et IR). Cela nous amène ainsi à considérer quatre sources d'informations L-VIS,G-VIS,L-IR,G-IR qui vont constituer les entrées de notre système de reconnaissance. Quant à la sortie, le système déterminera le type de l'obstacle : Piéton, Véhicule ou Fond d'image.
Les quatre sources d'informations considérées (L-VIS,G-VIS,L-IR,G-IR) sont de natures différentes. Les différents résultats exposés ont fait déjà preuve de complémentarités. D'un côté, il a été observé que la fusion des caractéristiques locales et globales a permis d'améliorer les performances de reconnaissance (voir section 3.4.1). De l'autre côté, le fait d'avoir des résultats de reconnaissance moins bons en VIS pour les piétons, mais meilleurs pour les véhicules, confirme la complémentarité entre les images VIS et IR.
Tout cela nous amène à conclure que la fusion des sources d'informations considérées apparait comme une solution prometteuse pour l'amélioration des performances de notre système de reconnaissance.

4.1.1 Fusion de caractéristiques ou de classifieurs ?

Dans les sections 2.3.1 et 2.3.2, nous avons étudié la typologie de la fusion des informations qui distingue entre deux niveaux de combinaison. La fusion de bas niveau, vise à combiner des caractéristiques issues directement de la source. À plus haut niveau, la fusion consiste à combiner les décisions (ici ce sont les résultats de classification). Ces deux niveaux sont implémentés dans notre système de reconnaissance où nous distinguons entre deux catégories de caractéristiques : locale (L) et globale (G) issues de deux modalités différentes (VIS et IR).
Il est difficile de faire un choix judicieux sans avoir étudié les sources d'informations à combiner. D'un côté, il est clair que la fusion de caractéristiques présente autant d'inconvénients que d'avantages si les sources à combiner sont de catégories différentes. De l'autre côté, la fusion des décisions issues de plusieurs classifieurs, bien qu'elle présente une certaine complexité de calcul, permet d'en cumuler des avantages. Ainsi, le fait que les sources d'informations considérées sont de catégo-

ries différentes, nous conduit à proposer de combiner des décisions de classifieurs spécialisés pour chaque type de caractéristiques. L'application de cette méthode est illustrée dans la figure 4.1. Ensuite, la meilleure règle de combinaison est choisie en comparant les performances de reconnaissance sur l'ensemble de validation. Deux types de combinaisons ont été envisagées :
- les combinaisons classiques en utilisant des opérateurs de combinaison de probabilité.
- les combinaisons basées sur la DST, afin de modéliser à la fois l'incertitude et l'imprécision des données.

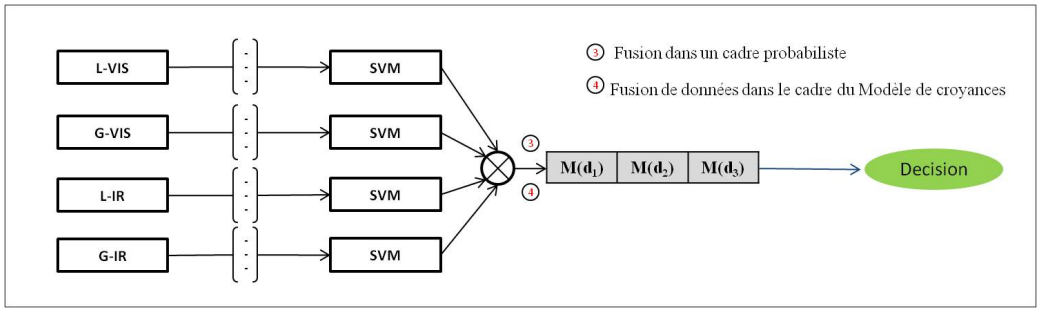


Figure 4.1. Les schémas de fusion envisageables

Il est évident que la solution proposée doit être pertinente et efficace, sa complexité est toute aussi importante. Cette complexité peut être réduite en analysant les différentes combinaisons de classifieurs.

4.1.2 Justification du choix de la DST pour la fusion de classifieurs

La décision du type de l'obstacle routier, revêt une importance cruciale pour un système d'aide à la conduite automobile. Ceci nous amène à implémenter les mécanismes nécessaires afin de prendre une décision précise. Toutefois, les décisions établies par des classifieurs ayant des niveaux de fiabilité différents, ne sont pas forcément parfaites, précises et certaines. Parmi les méthodes de fusion, seule la DST semble présenter un réel intérêt pour modéliser l'incertitude et l'imprécision. Ce choix est motivé pour plusieurs raisons. En premier lieu, la DST permet de prendre en compte et de modéliser à la fois l'imprécision, l'incertitude et l'incomplétude. En deuxième lieu, elle dispose, contrairement aux probabilités, d'outils spécifiques pour :

- formaliser la notion de conflit,
- décrire et gérer l'ignorance,

– fusionner les données par des règles de combinaison conçues sur mesure.

En vertu de l'ensemble de ces avantages, nous exploiterons la DST pour combiner les décisions de classifieurs. Les techniques de fusion fondées sur la probabilité serviront de comparatif.

4.2 Les outils de base des fonctions de croyance

Contrairement à la théorie des probabilités, le principe de la théorie des croyances repose sur la manipulation de fonctions définies sur des sous-ensembles et non seulement sur des singletons. Ces fonctions sont appelées *fonctions de masse*, ou encore masses de croyance.

4.2.1 Les fonctions de masse

Soit un ensemble fini $\mathcal{D} = \{d_1, d_2, .., d_n\}$, appelé cadre de discernement, correspondant à l'ensemble de toutes les décisions possibles (exhaustives et exclusives). Pour une problématique de classification, $\mathcal{D}$ regroupe l'ensemble de toutes les classes envisageables. L'espace des fonctions de masse m est donné par l'ensemble de toutes les combinaisons possibles des décisions 2^D. Un élément focal est un élément A de $2^{\mathcal{D}}$ tel que $m(A) > 0$. Il est important de mentionner qu'une fonction de masse respecte par définition les contraintes suivantes :

$$m(A) : 2^D \rightarrow [0, 1]$$
$$\sum_{A \subseteq D} m(A) = 1$$

Où la masse $m(A)$ représente le degré de croyance attribué à la proposition A. Les fonctions de masses permettent ainsi de modéliser des informations incertaines, de l'ignorance totale ($m(\mathcal{D}) = 1$) à la connaissance complète ($m(d_k) = 1$).

4.2.2 Autres représentations de croyance

Il existe d'autres manières de représenter les croyances que la fonction de masse :

Fonction de crédibilité	$bel(A)$	$=$	$\sum_{B\subseteq A, B\neq\emptyset} m(B)$
Fonction d'implicabilité	$b(A)$	$=$	$\sum_{B\subseteq A} m(B)$
Fonction de plausibilité	$pl(A)$	$=$	$\sum_{B\cap A} m(B)$
Fonction de communalité	$q(A)$	$=$	$\sum_{A\subseteq B} m(B)$

Ces différentes fonctions permettent de prendre en compte différents degrés de certitude ou de risque dans le processus de prise de décision.
En présence de sources d'informations jugées non totalement fiables, il est possible d'intégrer des coefficients d'affaiblissement dont les valeurs dépendent de la fiabilité de chacune des sources en question.

4.2.3 Affaiblissement des fonctions de masse

L'affaiblissement est un processus qui peut être opéré afin de prendre en compte de la fiabilité des sources dans le cadre des fonctions de croyance. Les premiers travaux sur l'affaiblissement ont été développés par Shafer [Sha78] puis formalisés par Smets [Sme93]. Ces travaux ont défini le processus de l'affaiblissement simple, qui a été généralisé par la suite par [Mer06, MQD08] pour être appelé *affaiblissement contextuel*.

4.2.3.1 Affaiblissement simple

L'affaiblissement simple consiste à réduire l'influence d'une source jugée non fiable. À partir d'une constante $\alpha \in [0, 1]$, appelée taux d'affaiblissement, l'opération d'affaiblissement de m est définie par :

$$\begin{aligned} {}^{\alpha}m(B) &= (1-\alpha)m(B), \forall B \subseteq \mathcal{D} \\ {}^{\alpha}m(\mathcal{D}) &= (1-\alpha)m(\mathcal{D}) + \alpha \end{aligned} \tag{4.1}$$

Plus α est grand, plus l'affaiblissement est important et moins la fonction de masse est engagée. L'affaiblissement simple agit de la même manière (avec le même coefficient) pour toutes les hypothèses issues de la même source. La généralisation de l'affaiblissement est basée sur l'idée que la fiabilité d'une source peut agir différemment sur les différentes hypothèses du cadre de discernement [Mer06, MQD08].

4.2.3.2 Affaiblissement contextuel

En utilisant l'opération d'affaiblissement contextuel, la dépendance au contexte de la fiabilité peut être prise en compte. Par le mécanisme général de la correction

proposée par [Mer06], différents états de fiabilité peuvent être modélisés. La généralisation de cette opération consiste premièrement à définir une partition notée $\Theta = \{\theta_1, .., \theta_G\}$ du cadre de discernement $\mathcal{D} = \{d_1, .., d_n\}$, deuxièmement, à quantifier des coefficients d'affaiblissement α_g, $g \in \{1, .., G\}$, sur chaque élément de la partition Θ. Chaque élément $\theta_G \in \Theta$ de la partition se voit donc affecter un taux d'affaiblissement α_g.
Disposant de G éléments θ_G dans la partition et donc de G taux d'affaiblissement rassemblés dans le vecteur $[\alpha_1, .., \alpha_g, .., \alpha_G]$, l'affaiblissement d'une fonction de masse m génère une distribution de masses de croyance (BBA) notée ${}^{\alpha}_{\Theta}m$ et définie par :

$$ {}^{\alpha}_{\Theta}m = m \cup m_1 \cup m_2 .. \cup m_g .. \cup m_G \tag{4.2} $$

où les BBA ${m_g}^D, g \in \{1, .., G\}$ sont définies par :

$$ \begin{aligned} m_g(\emptyset) &= 1 - \alpha_g \\ m_g(\theta_g) &= \alpha_g \end{aligned} \tag{4.3} $$

Dans la section suivante, nous présentons la méthodologie de la combinaison des fonctions de masses que nous avons adoptée dans notre système.

4.2.4 Règles de combinaison

Le formalisme des fonctions de croyance offre la possibilité de fusionner différentes sources d'informations, représentées par différentes fonctions de masses, à l'aide d'opérateurs appelés règles de combinaison. Plusieurs modes de combinaison ont été développés dans le cadre de la théorie des croyances, les deux principales combinaisons sont la combinaison conjonctive et disjonctive. Le choix peut se faire selon les propriétés désirées de l'opérateur (conjonctif, disjonctif, ou des compromis), selon son comportement dans des situations de conflit ou bien selon la complémentarité et les caractéristiques des sources d'informations.

4.2.4.1 Combinaison conjonctive

L'approche initiale a été introduite par Dempster [Dem67], puis elle a été reprise par Shafer [Dem67]. La combinaison conjonctive combine les fonctions de masse en considérant les intersections des éléments de 2^D.
Soient deux allocations de masse m_1 et m_2 issues de deux sources d'informations

fiables et distinctes. Ces deux fonctions peuvent être agrégées par un opérateur de combinaison conjonctif noté $\cap$. Le résultat de cette opération conduit à une fonction de croyance unique à laquelle correspond une fonction de masse définie par :

$$m_{1\cap 2}(A) = (m_1 \cap m_2)(A) = \sum_{B\cap C=A} m_1(B)m_2(C) \tag{4.4}$$

4.2.4.2 Combinaison disjonctive

À l'opposition de la combinaison conjonctive, la combinaison disjonctive [Sme93] est définie en considérant l'information complète fournie par les deux fonctions de masse. La combinaison de deux fonctions de masse m_1 et m_2, issues de deux sources d'informations dont l'une au moins est fiable, est donnée donc pour tout $A \in 2^D$ par :

$$m_{1\cup 2}(A) = (m_1 \cup m_2)(A) = \sum_{B\cup C=A} m_1(B)m_2(C) \tag{4.5}$$

Il est clair qu'alors $m(\varnothing) = 0$, en effet, par cette combinaison le conflit ne peut donc pas apparaître. En contrepartie, les éléments focaux de la fonction de masse résultante sont élargis, ce qui pousse à une perte de spécificité. Cette combinaison est peu ou pas employée car ce qui est recherché dans la plupart des applications est une fonction de masse plus focalisée. Une prise de décision sur les singletons ne sera possible que si les sources sont en accord sur ces singletons. Toutefois, cette approche est intéressante si nous ne savons pas modéliser les fiabilités des sources, leurs ambiguïtés et leurs imprécisions [Mar05].
Les règles conjonctives et disjonctives sont deux visions extrêmes du problème. D'autres règles adoptant un comportement intermédiaire ont été proposées [Yag87, DP92, Den08]. Récemment, Denoeux [Den08] a introduit une famille de règles dont nous ne présenterons que deux ; la règle conjonctive prudente et la règle disjonctive hardie, notées respectivement $\wedge$ et $\vee$. Ces règles permettent, contrairement aux règles conjonctives et disjonctives, de combiner des données issues de sources d'informations non distinctes. Une étude approfondie de toutes les règles présentées, ainsi que d'autres, est présentée dans [Kle08].

4.2.5 Prise de décision

La dernière étape de fusion consiste à prendre une décision sur la classe la plus vraisemblable. Contrairement à la théorie des probabilités, où le maximum a posteriori est le critère le plus souvent retenu, la théorie des croyances offre plusieurs règles de décision fondées sur la maximisation d'un critère :

- maximum de plausibilité,
- maximum de croyance,
- maximum de probabilité pignistique [Sme90],
- ou d'autres critères (pour plus de détails, [Mar05]).

Dans la plupart des applications, c'est le critère du maximum des probabilités pignistiques qui est souvent utilisé car il est basé sur un critère de compromis entre un maximum de plausibilité et un maximum de croyance. La probabilité pignistique est une mesure qui se base sur le principe de la raison insuffisante. Ce principe ne stipule qu'en l'absence de raison de privilégier une hypothèse plutôt qu'une autre, les hypothèses sont toutes supposées équiprobables. Ainsi, pour tout élément focal A de $m\odot(.)$ la masse sera redistribuée uniformément sur les éléments de A. Pour toute décision d_i, la probabilité pignistique est définie alors par :

$$Bet(d_i) = \sum_{A\subset 2^{\mathcal{D}}, d_i\in A} \frac{m(A)}{|A|(1-m(\emptyset))} \tag{4.6}$$

Bien que cette transformation offre une mesure subjective située entre la crédibilité et la plausibilité, elle est calculée uniquement sur les singletons du cadre de discernement. Une généralisation de cette transformation a été proposée par [BC09]. Elle sera détaillée dans la section 4.4.2.

4.3 Fusion de classifieurs dans le cadre de la DST

Ayant justifié théoriquement que la fusion de classifieurs dans le cadre de la DST est la meilleure solution pour fiabiliser un système de reconnaissance, nous détaillons dans cette section les différentes étapes nécessaires à sa mise en œuvre. Dans un premier temps, il est nécessaire de définir le cadre de discernement. Dans ce cas, le nombre d'hypothèses doit être assez limité car l'ensemble de toutes les combinaisons possibles des décisions croît exponentiellement avec celle de $\mathcal{D}$. Bien que l'exigence de temps réel soit fortement recommandée pour notre système, cela ne constitue pas un handicap car nous traitons seulement trois types d'OR : (Piéton, Voiture et Fond). Ce sont donc ces trois classes qui vont définir le cadre

de discernement. Ayant fixé le cadre de discernement $\mathcal{D}$, l'étape suivante consiste à construire des fonctions de masse spécifiques aux sorties des classifieurs.

4.3.1 Construction des fonctions de masse

Les fonctions de masses doivent être construites à partir des scores de classification fournis par des SVM. En effet, elles permettent de transformer la valeur de pertinence de la classification en une valeur dans $[0, 1]$. Cette étape nécessite une bonne connaissance du problème, puisque les fonctions de masse sont souvent constituées de manière empirique. Dans [BAC06], les auteurs définissent une fonction de masse typique au principe de classification SVM en exploitant la distance d'un vecteur caractéristique à l'hyperplan. Néanmoins, la certitude de classification ne peut pas être quantifiée par des algorithmes de classification non basés sur la notion de maximisation de la marge. Dans le cas général, deux étapes sont requises : l'estimation des probabilités à partir des sorties de classifieurs SVM, ensuite, la construction des fonctions de masse à partir des probabilités.

4.3.1.1 Transformation des sorties SVM en probabilité

Afin de transformer les sorties de SVM en probabilité, nous avons utilisé la méthode de [HT98] qui est basée sur une stratégie appelée Pairwise Coupling. Cette stratégie permet de combiner les sorties des classifieurs binaires, afin d'obtenir une estimation des probabilités a postériori $p_i = Prob(d_i/x)$ (Prob : probabilité) .
Soit $C_{i,j}$ le classifieur séparant les deux classes d_i et d_j et r_{ij} la probabilité résultante du classifieur $C_{i,j}$ tel que $r_{ij} = Prob(d_i/d_i, d_j)$. Il existe $|\mathcal{D}|(|\mathcal{D}| - 1)/2$ de variables μ_{ij} définies par $u_{ij} = p_i/(p_i + p_j)$. La valeur de p_i est calculée en utilisant une procédure itérative dont la condition d'arrêt est d'avoir une distance très faible entre r_{ij} et μ_{ij}. Cette distance $l(p)$ est mesurée de la façon suivante :

$$l(p) = \sum_{i<j} n_{ij}(r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}}) \tag{4.7}$$

Avec n_{ij} est le nombre d'exemples d'apprentissage qui appartiennent à $d_i \cup d_j$. Enfin, nous avons utilisé l'algorithme itératif 3, comme décrit dans [HT98], pour estimer p_i.

Algorithme 3 Estimation des probabilités a posteriori

1: Initialisation de p_i et calcul de μ_{ij}
2: **Répéter**
3: $\quad p_i \leftarrow p_i \frac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum_{j \neq i} n_{ij} \mu_{ij}}$
4: $\quad$ Re-normaliser p_i
5: $\quad$ Recalculer μ_{ij}
6: **Jusqu'à** convergence ($l(p)$ très faible)

4.3.1.2 Transformation des probabilités résultantes en fonctions de masse

Plusieurs méthodes de construction de modèles de fonctions de croyance ont été proposées dont deux catégories peuvent être ici distinguées. Une première catégorie de méthodes utilise des vraisemblances, estimées à partir d'un ensemble d'apprentissage, pour construire les masses de croyances [App91, Sme93, DPS01]. Une deuxième famille repose sur les distances entre observations et modèles [Den95]. Dans notre cas, nous nous intéressons à leurs définitions à partir d'un ensemble de vraisemblances. Plus particulièrement, nous avons adopté la méthodologie proposée par [DPS01] pour transformer les probabilités résultantes en fonctions de masse. Le principe de cette méthode consiste à interpréter les probabilités résultantes comme une distribution de probabilités subjectives qui sera ensuite convertie en une transformation consonante [1].

Par souci de simplicité de notations, nous supposons que les décisions de classifieurs d_i sont déjà ordonnées par des valeurs de probabilité décroissantes : $p(d_1) \geq \ldots \geq p(d_{|\mathcal{D}|})$. Une fonction de croyance consonante m correspondante à p se calcule de la façon suivante :

$$\begin{aligned} m\left(\left\{d_1, d_2, \ldots, d_{|\mathcal{D}|}\right\}\right) = m\left(\mathcal{D}\right) &= |\mathcal{D}| \times p(d_{|\mathcal{D}|}) \\ \forall\, i < |\mathcal{D}|,\ m\left(\{d_1, d_2, \ldots, d_i\}\right) &= i \times [p(d_i) - p(d_{i+1})] \\ m\left(.\right) &= 0 \qquad \text{sinon} \end{aligned} \tag{4.8}$$

Les fonctions de masses obtenues sont par la suite affaiblies en fonction de la fiabilité de chaque classifieur.

1. Une fonction de masse est dite consonante si et seulement si l'inclusion est une relation d'ordre totale sur ses éléments focaux : il $\exists$ une permutation sur les indices telle que : $D_{perm(1)} \subseteq D_{perm(2)} \subseteq D_{perm(n)}$)

4.3.2 Affaiblissement

Après avoir défini formellement, dans la section 4.2.3, la notion d'affaiblissement, nous décrivons l'application de ce processus dans notre système. L'affaiblissement simple a été implémenté afin de tenir compte de la fiabilité du classifieur. Ainsi, chaque masse m a été pondérée par un coefficient d'affaiblissement $(1-\alpha)$, avec α désignant le taux de reconnaissance du classifieur estimé sur un ensemble de validation.
Cette opération de correction agit de la même façon pour toutes les hypothèses issues de la même source. Dans le but d'attribuer une fiabilité spécifique à chaque hypothèse du cadre de discernement, nous avons implémenté aussi le processus d'affaiblissement contextuel. En effet, ce processus pourrait être très bénéfique dans notre système de reconnaissance qui est basé sur la fusion des modalités VIS et IR. Deux exemples à donner ; La reconnaissance des piétons est plus fiable dans les images IR dans la mesure où il est très contrasté par rapport au fond de l'image. En revanche, la reconnaissance des véhicules, surtout en mouvement, est plus facile en VIS qu'en IR. Ainsi, nous avons défini autant de facteurs d'affaiblissement contextuels que de nombre de classes dans le cadre de discernement.
Pour les deux types d'affaiblissement, les fiabilités ont été estimées à partir de la matrice de confusion en s'appuyant sur une base de validation.

4.3.3 Combinaison

La combinaison des fonctions de masse est une étape cruciale qui joue un rôle primordial dans l'obtention de bons résultats. Dans l'ensemble des règles présentées dans la section 4.2.4, nous représenterons dans le tableau 4.1 les quatre règles de combinaison considérées et leurs contextes d'utilisation.

Table 4.1. Les règles de combinaisons retenues et leurs contextes d'utilisation

$(m_1, m_2) \rightarrow m_{1\odot 2}$	Sources fiables	Sources non fiables
Sources distinctes	$m_{1\cap 2}$ (conjonctive)	$m_{1\cup 2}$ (disjonctive)
Sources non distinctes	$m_{1\wedge 2}$ (conjonctive prudente)	$m_{1\vee 2}$ (disjonctive hardie)

4.3.4 Prise de décision

Dans le cadre du MCT (modèle des croyances transférables), il est intéressant d'utiliser les fonctions de croyance pour gérer des données imprécises et mettre au

point des processus adéquats de fusion. En revanche pour prendre une décision, les probabilités sont mieux adaptées, car à ce stade l'imprécision ne fait que bruiter le traitement [Kle08]. C'est pourquoi nous avons utilisé le critère du maximum de probabilité pignistique. Ainsi, pour une observation x la décision prise (d_i) consiste à choisir l'élément k possédant la plus grande probabilité pignistique $Bet(d_i)(x)$ (Eq 4.9).

$$Bet(d_i)(x) = max_{i \leq k \leq n} Bet(d_k)(x) \tag{4.9}$$

4.4 Proposition d'une stratégie de classification à deux niveaux de décision

Le système de reconnaissance d'OR doit réaliser un compromis entre la précision et le temps de calcul, compatible avec les contraintes d'implantation d'un système embarqué. Il est certes qu'en utilisant la DST, la précision peut tirer profit de la diversité et de la complémentarité des différentes sources d'informations. Néanmoins, le temps de calcul et la complexité algorithmique seront d'autant plus importants que le nombre de sources d'informations. De plus, l'utilisation d'une source d'information non fiable dans le processus de fusion peut mener à des résultats erronés. Cela nous a amené à proposer une stratégie de fusion qui consiste, en premier lieu, à sélectionner les meilleures sources d'informations à combiner. En deuxième lieu, cette stratégie comporte deux niveaux de décision permettant d'accélérer grandement le processus de prise de décision tout en tirant profit de l'utilisation de la DST. Cette stratégie est illustrée dans la figure 4.2, elle comporte deux niveaux de décision dont le deuxième est optionnel et ne sera exploité que si la réponse du premier classifieur est incertaine. Afin que cette stratégie soit efficace, trois conditions nécessaires doivent être satisfaites :

1. Le premier niveau de décision doit être occupé par un classifieur plus performant que celui du deuxième niveau.
2. Les résultats de fusion doivent être suffisamment satisfaisants. Ce qui revient à dire que les sources à combiner doivent être pertinentes et suffisamment complémentaires.
3. La méthode de mesure de certitude doit être suffisamment précise pour pouvoir évaluer la précision de la décision prise par le premier classifieur.

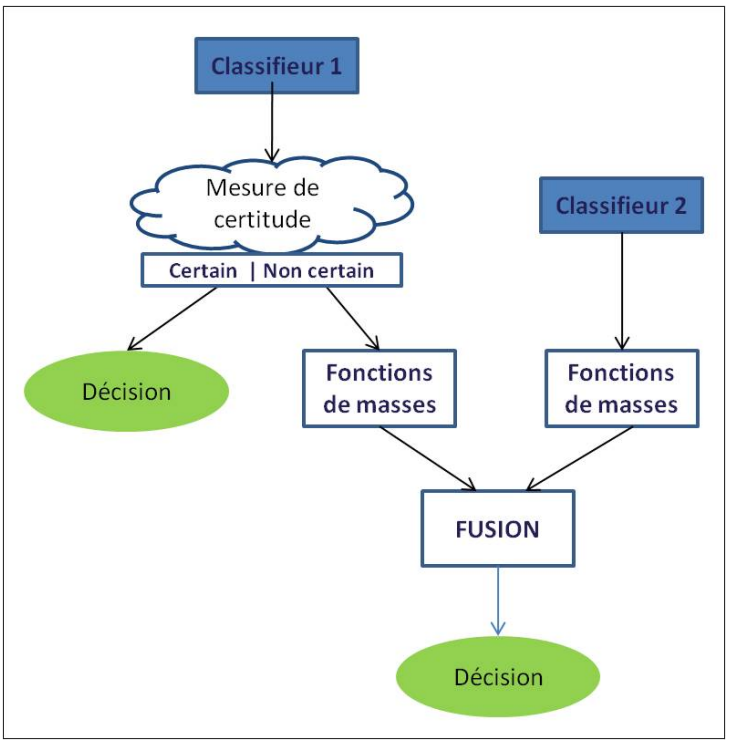


Figure 4.2. Stratégie de classification à deux niveaux

Pour mesurer la certitude du résultat de classification, nous proposons d'utiliser deux méthodes :

- la comparaison de la largeur de la marge et la distance à l'hyperplan,
- l'utilisation de la transformation pignistique généralisée [BC09].

4.4.1 Distance à l'hyperplan

Comme nous l'avons signalé dans la section 3, lors du processus d'apprentissage, le classifieur SVM cherche à déterminer un hyperplan optimal dont la distance minimale aux exemples d'apprentissage est maximale. Pour classifier un exemple, SVM fournit, à part l'étiquette, une distance à l'hyperplan, qui a été construit en apprentissage (figure 4.3). Nous proposons d'explorer cette information afin d'évaluer la confiance attribuée à la classification. En fait, si la distance entre un point de test et l'hyperplan est supérieure à la marge, la décision prise par SVM sera considérée comme décision finale. Sinon, nous utiliserons une deuxième source d'information afin de fiabiliser la prise de décision.

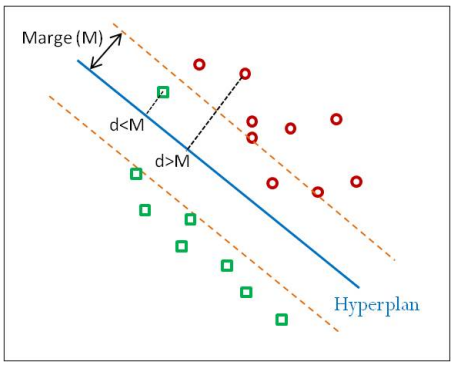


Figure 4.3. Notion de marge en SVM et distance entre un vecteur et l'hyperplan

4.4.2 La transformée pignistique généralisée

La transformation pignistique généralisée (TPG) [BC09] est, comme l'indique son nom, une généralisation de la transformation pignistique. En effet, elle inclut en particulier un paramètre γ. Dans le cas où $\gamma = 1$, nous retrouvons la même équation de la transformée pignistique originale (Eq 4.6). Quand $\gamma > 1$, la généralisation de la transformation pignistique entraîne une décision qui ne sera pas forcément focalisée sur une seule éventualité, mais sur plusieurs (au maximum γ).

$$B_\gamma(B) = m(B) + \sum_{B \subset A \subset \mathcal{D}} \frac{m(A).|B|}{N(|A|,\gamma)}, \forall B \subseteq \Delta_\gamma \tag{4.10}$$

Avec :

$$N(|A|,\gamma) = \sum_{k=1}^{\mathcal{D}} \frac{|A|!}{k!(|A|-k)!}.k \tag{4.11}$$

Avec B_γ désignant le résultat de la transformée de m, $|A|$ le cardinal de A et Δ_γ l'ensemble des éléments de cardinalité inférieure ou égale à γ. Nous proposons de prendre $\gamma = 2$ (deux niveaux de décision) et de se baser sur cette cardinalité afin d'évaluer la précision de classification fournie par le premier classifieur. Si la solution retournée est un singleton, le classificateur fournit une décision sûre, ainsi la classe proposée va être considérée comme une décision finale. Dans le cas où $|A| = 2$, nous proposons d'utiliser un deuxième niveau de classification afin de prendre une décision parmi les deux classes proposées. Par exemple, si la solution retournée par le TPG est $\{P, V\}$, nous utilisons un deuxième niveau de classificateur en utilisant un SVM qui permet de séparer entre les deux classes : Piétons et Véhicule.

Ayant souligné que les performances de la stratégie proposée dépendent de la méthode de mesure de certitude utilisée, les deux méthodes (distance à l'hyperplan et TPG) ont été implémentées et comparées.

4.5 Evaluation des performances de la reconnaissance des obstacles routiers

Nous présentons dans cette section les différentes expérimentations que nous avons réalisées sur la même base d'images muti-classe décrites dans la section 3.4 (*Tetra6*). L'ensemble des données comprend un total de 986 objets annotés en VIS et en IR répartis entre des Piétons, Véhicules et Fond d'images. La Figure 4.4 donne un exemple d'objets corrélés en VIS et IR.

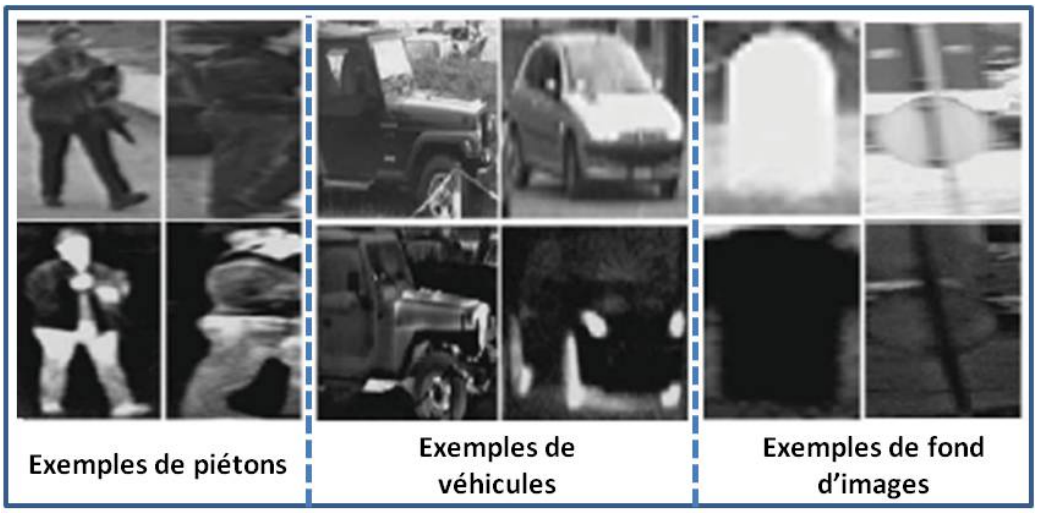


Figure 4.4. Quelques exemples d'objets contenus dans la base de test VIS et IR

Avant de présenter l'apport de la fusion, il est important de donner dans le tableau 4.2 les résultats de la classification obtenus pour chaque source d'information : L-VIS,G-VIS,L-IR,G-IR.

Table 4.2. Taux de classification obtenus pour chacune des caractéristiques considérées

Caractéristiques	L-IR	G-IR	L-VIS	G-VIS
Taux de classification	63.94%	87.41%	55.11%	64.62%

Le tableau 4.2 résume les taux de classification moyens obtenus par chacune des caractéristiques en utilisant un SVM multiclasses. Nous voyons que la précision maximale obtenue n'est pas totalement satisfaisante (autour de 87.5%). Afin

d'améliorer les performances de notre système, nous proposons de fusionner ces informations.
Dans le but d'améliorer les performances de notre système, nous avons choisi de combiner les sorties SVM/SURF. Comme nous l'avons signalé dans la section 4.3, la fusion de classifieurs dans le cadre de la DST nécessite d'ajuster différents paramètres. En effet, les performances du module de fusion dépendront de :

- la sélection des sources d'informations (fiabilité et complémentarité),
- l'affaiblissement des fonctions de masse,
- la stratégie de classification (utilisation d'un seul ou de deux niveaux de classification).

Dans la suite, nous examinons l'impact de chaque paramètre.

4.5.1 Analyse des paramètres de fusion

Dans le tableau 4.3, nous présentons une comparaison entre les différentes méthodes de fusion pour toutes les combinaisons possibles des sources d'informations. Pour chaque combinaison, nous comparons les résultats aux meilleurs résultats obtenus par les méthodes classiques de combinaison probabiliste (MCP) : somme, somme pondérée, produit et produit pondéré (les formules sont données dans 2.3.2.2). De même pour l'application de la DST, nous présentons les meilleurs résultats obtenus selon la méthode d'affaiblissement exercée (S : simple, C : contextuel, N : pas d'affaiblissement). Il est important de mentionner que les taux d'affaiblissement simple utilisés sont les taux de reconnaissance spécifiques à chaque modalité. Quant à l'affaiblissement contextuel, les taux d'affaiblissement sont les taux de reconnaissance de chaque classifieur pour chaque hypothèse (P,V,F). Par exemple, pour G-IR⊗L-VIS les taux d'affaiblissement utilisés sont :

- 2 taux d'affaiblissement simple estimés à partir des taux de reconnaissance dans les modalités G-IR et L-VIS
- 6 taux d'affaiblissement contextuel estimés à partir des taux de reconnaissance dans les modalités G-IR et L-VIS pour chaque hypothèse P, V et F.

Ces différents taux ont été estimés sur la base de validation.
Afin d'évaluer l'impact de la typologie de fusion, nous comparons les résultats selon les deux typologies : fusion de classifieurs et fusion de caractéristiques. Ainsi, nous confrontons les résultats obtenus avec ceux obtenus avec l'approche par fusion des caractéristiques (FC).

Les combinaisons basées sur la DST permettent de modéliser à la fois l'incertitude et l'imprécision des données. Cette propriété semble très importante pour

Table 4.3. Résultats obtenus en utilisant différentes méthodes de fusion

Caractéristiques	DST		MCP	FC
	MA	TR	TR	TR
G-IR, L-IR	S	88.09	86.05	89.8
G-VIS, L-VIS	N	72.78	70.40	71.08
G-IR, L-VIS	S	**90.81**	88.09	87.41
L-IR, G-VIS	S	78.57	78.23	73.12
G-IR, G-VIS	S	87.41	85.37	80.27
L-IR, L-VIS	N	74.48	72.44	70.06
G-IR, L-IR, G-VIS	N	88.43	87.75	80.27
L-IR, G-VIS,L-VIS	C	77.55	77.55	74.14
L-IR, G-IR,L-VIS	S	86.56	84.69	80.27
L-VIS, G-IR,G-VIS	N	88.43	80.27	84.69
L-VIS, G-VIS, L-IR,G-IR	C	90.47	80.27	88.43

notre cadre d'application. En effet, les techniques de fusion utilisées basées sur la DST ont dépassé les résultats obtenus par les méthodes probabilistes. En ce qui concerne la typologie de fusion, on remarque que la fusion évidentielle des classifieurs permet d'obtenir des résultats globalement meilleurs comparativement à l'utilisation de l'approche par fusion de caractéristiques. Toutefois, cette remarque n'est pas vérifiée pour toutes les combinaisons, à l'occurrence de L-VIS$\otimes$G-VIS et de L-IR$\otimes$G-IR. Cela montre que la fusion DST est surtout intéressante quand il s'agit de combiner des sources d'informations issues de différents capteurs.
On voit d'après les résultats du tableau 4.3 que les meilleurs résultats sont obtenus en combinant L-VIS et G-IR. Ce résultat est intéressant dans la mesure où le système présente de bons résultats en combinant seulement deux sources d'informations. Ce résultat s'explique par le fait que les deux sources L-VIS et G-IR sont indépendantes et surtout complémentaires. D'un côté, le L-VIS permet de caractériser les apparences locales en VIS. De l'autre, la modalité G-IR apporte des informations complémentaires de forme et de texture. Cette complémentarité peut également être confirmée par l'analyse des résultats présentés dans le tableau 4.4.

Table 4.4. Résultats obtenus en fonction des règles de combinaison des fonctions de masse

Règle de combinaison	$\cap$	$\cup$	$\wedge$	$\vee$
G-IR, L-IR	87.41	76.87	88.09	84.35
G-IR, L-VIS	**90.81**	73.80	89.79	89.79

Le tableau 4.4 montre que la règle optimale de combinaison L-VIS⊗G-IR, est la conjonctive. Ce résultat est prévisible vu que les sources sont indépendantes. Cependant, cela n'est pas vérifié pour les sources (G-IR, L-IR), où la règle optimale de combinaison est la prudente. Ceci semble logique car ces deux sources sont issues du même capteur (caméra IR), et sont ainsi légèrement dépendantes.
La conclusion majeure que nous pouvons tirer des résultats présentés ci-dessus est que la combinaison des 4 sources d'informations s'avèrent inutile puisque les meilleurs résultats ont été obtenus en ne combinant que les sources (G-IR, L-VIS). Cela permettra certainement d'optimiser la complexité globale du système de reconnaissance. Pour continuer dans cette voie, il est important d'évaluer la stratégie de classification à deux niveaux que nous avons proposée dans la section 4.4.

4.5.2 La stratégie de classification à deux niveaux

La stratégie de classification à deux niveaux que nous avons proposé comporte deux niveaux de décision dont le deuxième est optionnel et ne sera exploité que dans le cas d'incertitude portant sur la réponse du premier classifieur. Dans ce qui précède, nous avons sélectionné les meilleures sources d'informations à combiner qui sont les caractéristiques : L-VIS et G-IR. L'application de la stratégie de classification proposée consiste à placer le classifieur de G-IR dans le premier niveau puisque c'est le classifieur le plus performant. Quant au deuxième niveau de classification, il sera occupé par le classifieur sur L-VIS qui est moins performant que le premier.
Dans le tableau 4.5, nous évaluons les résultats obtenus par l'application de cette stratégie. De plus, nous mettons en évidence les méthodes de mesure de certitude à savoir la distance à l'hyperplan (SVM-DH) et la transformation pignistique généralisée (TPG). Concernant ces résultats, nous présentons les meilleurs en fonction de la méthode d'affaiblissement exercée (MA). Etant donné que le G-IR est le seul classifieur fort, nous donnons les résultats des différentes combinaisons en plaçant ce classifieur en premier niveau.

À partir des résultats présentés dans le tableau 4.5, on peut constater que la stratégie de classification proposée a permis non seulement de réduire la complexité du système, mais aussi d'améliorer légèrement les performances de reconnaissance. L'amélioration est enregistrée surtout pour G-IR⊗G-VIS et G-IR⊗L-IR. Nous expliquons cela par le fait que le deuxième niveau de classification implique un classifieur moins performant qui peut exercer un impact négatif sur le processus de

Table 4.5. Résultats obtenus en utilisant la stratégie de classification à deux niveaux

Caractéristiques	Fusion évidentielle	La stratégie proposée			
		TPG		SVM-DH	
		MA	Précision	MA	Précision
G-IR,G-VIS	87.41	C	88.43	C	88.09
G-IR,L-VIS	**90.81**	C	90.81	S	90.81
G-IR,L-IR	88.09	C	89.11	N	88.09

fusion. Cet impact est particulièrement marqué en cas de mauvaises classifications avec des taux importants de certitude. En d'autres termes, si le classifieur est sûr de sa décision, malgré qu'il se trompe.

Il est fort de constater que pour G-IR⊗L-VIS, le second niveau de classification n'a été utilisé que pour 16,32% du nombre total d'exemples contenus dans la base du test. Cela confirme le gain considérable en temps de calcul, obtenu en utilisant la stratégie proposée.

En ce qui concerne l'affaiblissement des fonctions de masse, la comparaison des résultats obtenus dans les tableaux 4.3 et 4.5 permet de révéler l'importance de l'affaiblissement contextuel. Cette méthode permet de réduire l'impact négatif de la précision du classificateur en atténuant le poids des fausses décisions.

La mesure de certitude liée à la décision du premier classifieur est une condition nécessaire pour garantir une prise de décision fiable. La comparaison présentée dans le tableau 4.5 montre que les résultats obtenus en se basant sur la mesure de TPG sont légèrement meilleurs que ceux basés sur la distance à l'hyperplan. Ce résultat semble logique puisque la mesure de certitude SVM-DH est moins précise que la TPG. En effet, elle accorde la même confiance pour les instances se trouvant légèrement ou beaucoup plus distants de la marge. En revanche, la TPG fournit une solution plus précise, en transformant les sorties de SVM en probabilités.

Enfin, nous pouvons conclure que la stratégie proposée de la classification à deux niveau convient parfaitement à notre problématique de reconnaissance des obstacles routiers. Elle permet de réduire la complexité du problème et d'accélérer le temps de calcul sans compromettre les performances du système. Après avoir évalué tous les paramètres, nous présentons dans la figure 4.5 le schéma final de cette stratégie.

4.5.3 Discussions

Plusieurs conclusions peuvent être tirées des résultats expérimentaux présentés. Les résultats obtenus avec la fusion fondée sur la DST montrent une réelle amélio-

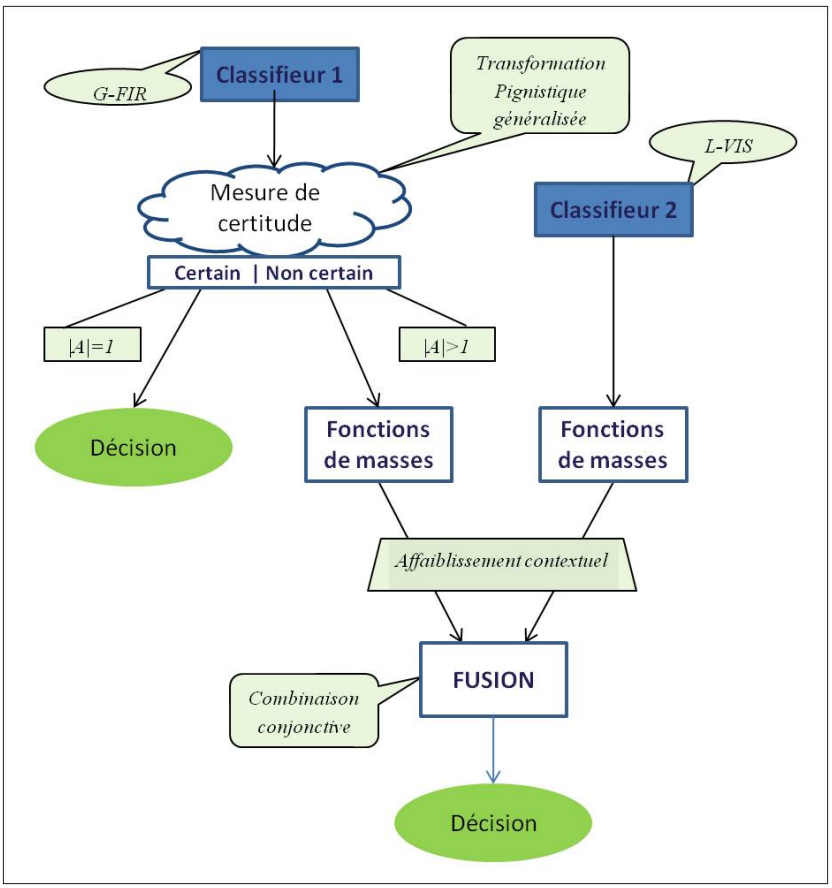


Figure 4.5. Descriptif de la stratégie de décision basée sur la classification à deux niveaux

ration par rapport aux performances individuelles des classifieurs (quelle que soit la modalité et l'ensemble des caractéristiques). De plus, la comparaison avec les méthodes de fusion probabilistes ou par la fusion de caractéristiques confirme l'intérêt de la fusion évidentielle. Bien que le gain en termes de performances semble modeste, toute amélioration dans le cadre de l'assistance à la conduite est bénéfique.

De plus, il a été constaté que les performances de reconnaissance ne s'améliorent pas proportionnellement avec le nombre de sources à combiner. La sélection de sources doit se focaliser plutôt sur une analyse fine de leur complémentarité. L'analyse des différentes combinaisons de caractéristiques nous a permis de déterminer les meilleures sources à combiner qui sont L-VIS⊗G-IR. Ces caractéristiques sont complémentaires et se distinguent par la modalité et la manière dont elles sont extraites.

Afin de réduire la complexité du problème, nous avons proposé une stratégie ba-

sée sur deux niveaux de classification. Cette solution permet d'accélérer le temps de traitement vu que l'utilisation du deuxième classificateur intervient seulement en cas d'incertitude liée à la décision du premier classifieur. Les expérimentations réalisées mettent en valeur le fait que cette stratégie permettant de réduire la complexité du système sans compromettre les performances de reconnaissance.

4.5.4 Bilan

Les résultats présentés dans ce chapitre prouvent que le système de reconnaissance basé sur la fusion évidentielle est capable de traiter des tâches compliquées de reconnaissance. Les deux sources d'informations sélectionnées (G-IR et L-VIS) sont basées sur SURF/SVM et permettent de caractériser les textures, les formes en IR et les apparences locales d'objets en VIS. Enfin, le module de fusion basé sur la DST permet de combiner les données provenant de ces deux sources de façon robuste en couvrant les problèmes d'imprécision, d'incertitude et de conflits.
Bien que les différentes approches proposées ont montré des performances similaires, nos expériences permettent de conclure la supériorité des méthodes de fusion basées sur la DST par rapport aux méthodes probabiliste de fusion. Néanmoins, cet avantage n'est pas significatif quand il s'agit de combiner des sources unimodales (exemple : G-IR et L-IR). Il apparaît en effet, dans les résultats présentés précédemment, que la fusion des caractéristiques G-IR⊗L-IR donne un résultat meilleur que l'approche DST (FC : 89.8%, DST : 88.09%). Ceci peut être expliqué par le fait que l'approche par fusion de caractéristiques peut donner des résultats intéressants quand il s'agit de combiner des sources homogènes (issues de la même modalité). Tandis que, le meilleur résultat obtenu avec la DST (G-IR⊗L-VIS) est de 90.81%, mais c'est la conséquence de la fusion multimodale des capteurs IR et VIS. Dans le tableau 4.6, nous comparons les taux de F-score spécifiques à la reconnaissance de chaque objet routier.

Table 4.6. Comparaison entre les meilleurs taux de F-score obtenus avec une fusion unimodale et multimodale

Sources	Typologie de fusion	Taux de F-score		
		Piéton	Véhicule	Fond
G-IR⊗L-IR	Fusion de caractéristiques (1 classifieur SVM)	**93.7**	85.4	87.8
G-IR⊗L-VIS	Fusion basée sur la DST (2 classifieurs SVM)	94	89.1	89.1

Les comparaisons données dans le tableau 4.6, ne sont significatives que pour la

classe Véhicule où on remarque une amélioration autour de 4%. En ce qui concerne la classe Piéton, l'amélioration est faible ce qui laisse subsister le doute quant au coût de mise en œuvre du module de fusion multimodal (prix, calibrage, temps de calcul). Quoi qu'il en soit, les deux résultats sont satisfaisants mais pas assez robustes pour assurer l'implantation d'un système de détection générique. Ainsi, ces résultats nous conduisent à envisager l'implantation d'un seul capteur IR pour la mise en œuvre d'un système embarqué de détection de piétons. Cette partie est développée dans le chapitre suivant.
Finalement, nous pensons que le formalisme de fusion fondé sur les fonctions de croyance, même s'il ne résout pas précisément le problème de reconnaissance générique d'OR, a tout de même pu permettre d'améliorer les taux de reconnaissances par rapport aux méthodes de fusion probabilistes plus classiques. En outre, les outils de base de la DST nous ont permis d'analyser l'indépendance et la complémentarité non seulement au niveau capteur mais aussi au niveau caractéristiques.

4.6 Conclusion

Dans ce chapitre, nous avons montré que la fusion de données constitue un moyen pour l'amélioration des performances d'un système de reconnaissance d'obstacles routiers. Elle permet en effet de fiabiliser les données et d'apporter une représentation plus riche de la scène routière. Après avoir présenté le contexte d'utilisation, nous avons mentionné les différents niveaux, stratégies et cadres de fusion retenus dans notre système. Pour réduire la complexité du système et dans le but de respecter la contrainte de temps réel, une stratégie de classification à deux niveaux de décision a été proposée. Les résultats expérimentaux obtenus montrent l'intérêt pratique de la fusion de modalités visible et infrarouge pour la reconnaissance des obstacles routiers. Néanmoins, le coût des capteurs ainsi que la complexité du calcul en limitent l'application. Ces considérations ainsi que les conclusions tirées des résultats obtenus, nous incitent à explorer uniquement la modalité infrarouge pour une application de détection de piéton. Cette idée est approfondie dans le chapitre suivant où nous proposons un système rapide de détection de piétons en IR.

Chapitre 5

Application à la détection de piétons en infrarouge lointain

Introduction

Les chapitres précédents nous ont permis de justifier le choix de capteurs et de proposer des techniques de représentation et de classification ayant pour but d'apporter une réponse fiable au problème de détection des obstacles routiers. Dans le présent chapitre, nous mettons en œuvre ces techniques dans un cadre applicatif lié à la mise en place d'un système de détection de piéton. Ce système a été validé par différentes expérimentations sur des images et des séquences routières en milieu urbain. Ces expérimentations montrent que le système proposé produit des résultats précis et robustes face aux problèmes de changements d'échelle et d'occultations partielles de piétons.

5.1 Préliminaires

La détection et le suivi de piétons ont fait l'objet de nombreux travaux pour une multitude d'applications. Dans le cadre de l'aide à la conduite automobile, ce sujet est particuliairement difficile et important car le piéton est l'objet le plus vulnérable présent dans une scène routière.
À l'issue de l'étude bibliographique, rapportée dans le premier chapitre, nous avons constaté que la problématique de détection et de suivi d'obstacles dans des scènes dynamiques, ne peut être résolue convenablement sans recourir à des techniques de reconnaissance de catégories d'objets dans les images. Ainsi, les chapitres 2,3 et 4 se sont attachés à examiner l'état de l'art et à justifier les choix techniques en ce qui concerne les techniques de reconnaissance d'objets. Dans ce chapitre, nous

mettons à profit les conclusions tirées des expérimentations menées à la mise en place d'un système de détection de piéton.

5.1.1 Synthèse des résultats de reconnaissance

Pour répondre à la problématique de reconnaissance d'objets dans les images, nous avons centré notre réflexion autour de trois axes : la représentation, la classification et la fusion d'informations. Ainsi une évaluation expérimentale a été réalisée autour de ces trois axes. La remarque la plus importante que nous avons tirée de tous les résultats présentés est que le système donne des résultats très satisfaisants quant à la reconnaissance de piétons à partir des images IR. Afin de justifier la pertinence de la méthode de la représentation et de la classification, les résultats ont été confrontés avec d'autres modèles de représentation et d'autres fonctions noyaux pour la classification par un SVM. En ce qui concerne la fusion d'informations, son apport a été surtout remarquable au niveau de la fusion de caractéristiques globales et locales.
Par conséquent, ces interprétations nous ont amenés à intégrer les éléments développés dans un système de détection et de suivi de piéton en IR. L'ensemble des caractéristiques locales et globales extraites des ROI seront évaluées par SVM afin de valider la présence d'un piéton. Il reste à concevoir une méthode permettant de générer des ROI susceptibles de contenir des piétons, ce qui n'est pas une tâche aisée.

5.1.2 Schéma de l'application

La détection automatique d'objets dans une image, uniquement par extraction des primitives visuelles, est une tâche très complexe. En effet, le processus de détection est autant, voir plus compliqué que celui de la reconnaissance. Toutefois, le fait d'avoir validé un classifieur spécialisé dans la discrimination des piétons, nous permettrait de résoudre une partie de la problématique. Dès lors, un balayage exhaustif de l'image avec plusieurs valeurs d'échelles permettrait de résoudre l'autre partie du problème. Mais, comme nous l'avons signalé, cette méthode est très coûteuse en temps de calcul. Ce que nous proposons est d'exploiter les caractéristiques des images IR et de chercher dans les zones claires des indices caractéristiques des piétons. De ce fait, nous avons procédé à l'extraction des points d'intérêt SURF [BTG06] dans des zones à fort contraste. La mise en correspondance des descripteurs SURF avec le Vocabulaire Visuel Hiérarchique permettra d'évaluer l'appartenance du motif (description) à la catégorie piéton en faisant voter les

clusters. Néanmoins, ce vote ne fournit qu'une mesure de similarité d'apparence. Ainsi, nous proposons d'injecter une information spatiale implicite au sein du Vocabulaire Visuel permettant d'indexer la position relative du piéton par rapport à l'emplacement du POI. Nous nous sommes également focalisés sur la résolution du problème d'occultation partielle qui était perçu à l'origine de l'utilisation de l'IR-lointain. Afin de surmonter les difficultés de séparation entre les piétons occultés, nous proposons de commencer par une phase de détection des têtes. Ces régions peuvent être identifiées en utilisant un Vocabulaire Visuel Hiérarchique regroupant les caractéristiques locales de ces régions spécifiques. Ensuite, le système de détection proposé procède par la construction des fenêtres d'intérêt qui sont par la suite validées par SVM. Finalement, les piétons détectés feront l'objet d'un processus de suivi basé sur l'appariement temporel des descripteurs SURF. La figure 5.1 illustre le schéma global du système de détection proposé.

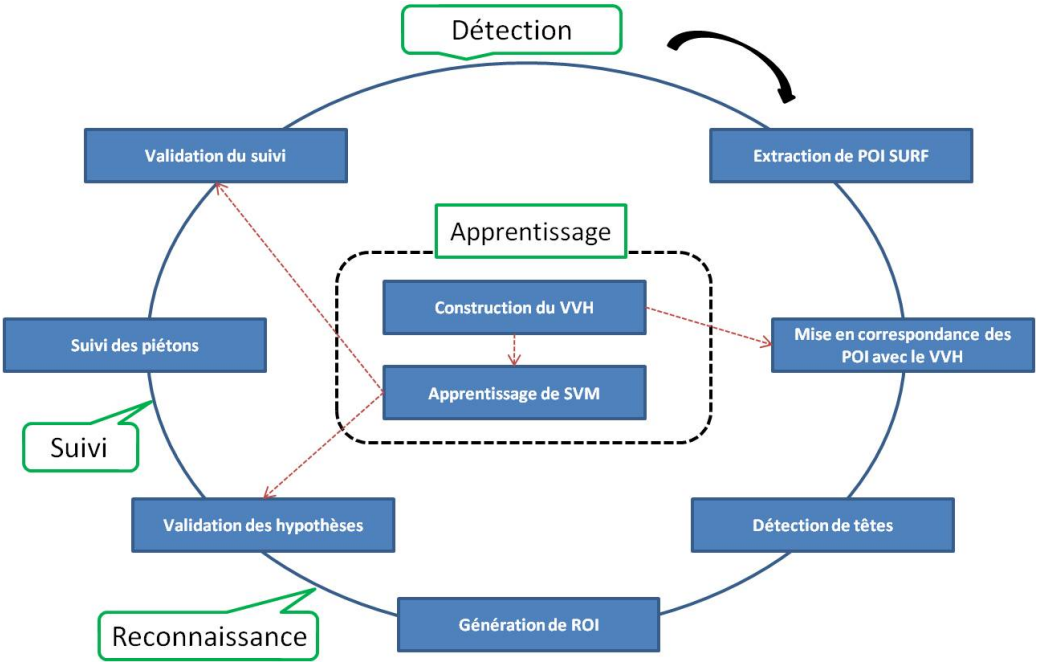


Figure 5.1. Le schéma global du système de détection et de suivi proposé

La figure 5.1 expose un schéma illustrant les processus clés du système proposé. Nous distinguons trois étapes principales : l'apprentissage, la détection et le suivi. La phase d'apprentissage fait appel à deux processus : la construction du VVH et l'apprentissage d'un classifieur SVM. Sur la figure, les flèches en rouge reliant les processus sont censées indiquer à quel moment le VVH et le SVM sont utilisés.

5.2 Le système de détection et de suivi proposé

Dans cette section, nous décrivons les processus principaux intervenant dans notre système de détection de piétons.

5.2.1 Apprentissage

Notre système de détection s'appuie sur une phase d'apprentissage durant laquelle deux tâches sont réalisées : la construction du VVH et l'apprentissage de SVM. Cette phase se distingue de celle opérée pour l'apprentissage du système de reconnaissance de piéton, décrite dans le chapitre 3. En effet, le contenu du VVH est modifié : d'uns part, le VVH a été construit après avoir rassemblé et regroupé les descripteurs SURF localisés dans des régions de tête. D'autre part, une information spatiale implicite a été injectée au sein de chaque cluster du VVH permettant d'encadrer les éventuelles régions contenant des têtes. En effet, nous proposons d'associer un paramètre (r) représentant le ratio entre l'échelle (ρ) à laquelle le POI a été extrait, et la distance (d) à la plus proche bordure de la fenêtre qui englobe le piéton (figure 5.2). L'enregistrement de ce paramètre permettra, dans la phase de détection, de générer une fenêtre autour d'un POI ayant un descripteur similaire (ayant des caractéristiques locales proches). La figure 5.2 illustre le principe de cette méthode.

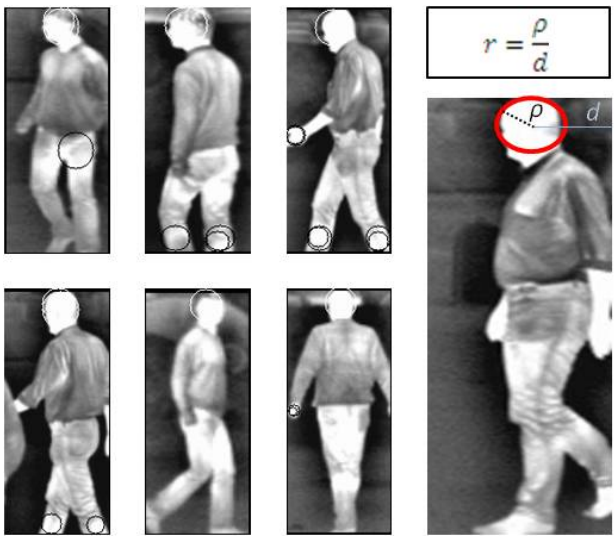


Figure 5.2. Extraction de POI SURF localisés dans des régions de têtes (cercles en blanc) et l'enregistrement du rapport entre l'échelle et la distance à la plus proche bordure

Après la phase d'apprentissage, le Vocabulaire Visuel construit est représenté

par un arbre de clusters dont chacun est caractérisé par :

- C_i : Un centroïde (vecteur descripteur moyen des points SURF).
- R_i : Un rayon du cluster (distance Euclidienne du centre au descripteur le plus éloigné).
- r_i : Une valeur moyenne des ratio r associés au cluster.

Dans la suite nous expliquons comment ces paramètres vont être utilisés afin de parvenir à détecter et à localiser des piétons.

5.2.2 Génération des hypothèses

Le processus de génération d'hypothèses sur la présence possible de piétons s'appuie sur une étape fondamentale de mise en correspondance des POI SURF extraits d'une image de test avec le VVH. À la fin de cette étape, des fenêtres d'analyse sont définies autour des éventuelles positions de têtes. Ensuite, ces fenêtres sont traitées afin de parvenir à définir des ROI qui englobent les totalités des corps des piétons.

5.2.2.1 Détection de têtes et génération de ROI

Le processus de détection de têtes est accéléré par l'exploitation de la représentation hiérarchique du VV. Comme dans l'étape de reconnaissance, une exploration partielle de l'arbre est généralement suffisante (voir section 3.2.3). Nous nous proposons simplement ici de rappeler les grandes lignes. Lors de la mise en correspondance, il n'est pas nécessaire d'examiner les sous-arbres dont le nœud père n'a pas été activé. Un nœud s'active si la distance Euclidienne entre le descripteur du POI et le centroïde du cluster désigné est inférieure à son rayon. Par conséquent, la mise en correspondance est réalisée en appliquant un simple algorithme itératif de parcours en profondeur du VVH. À la différence du processus d'extraction de caractéristiques (voir section 3.2.3), nous ne considérons que les votes des nœuds les plus profonds dans l'arbre. Parmi ces votes, seulement celui qui maximise la valeur de l'activation du cluster $A_{i,k}$ par le POI considéré, est retenu. Ainsi, la mise en correspondance de chaque POI extrait de l'image de test, génèrera une seule fenêtre d'intérêt.
La fenêtre d'intérêt est déterminée en fonction du paramètre r_i associé au cluster i, maximisant $A_{i,k}$, et l'échelle du POI ρ_k. Cette fenêtre est centrée sur la position du POI et nous y proposons les dimensions suivantes :

$$\begin{aligned} L &= 2 \times \frac{\rho_k}{r_i} \\ H &= L/2 \end{aligned} \tag{5.1}$$

Cette fenêtre permet ainsi de définir une zone d'intérêt susceptible de contenir une tête. Bien que la hauteur de la fenêtre H soit définie d'une manière grossière, sa valeur sera réajustée après la détermination de la région d'intérêt qui englobe entièrement le piéton. En ce qui concerne la valeur de L, la formule proposée est justifiée par le fait que r_i permet de retrouver la distance à la bordure de la fenêtre qui englobe le piéton en utilisant le paramètre r_i du cluster activé.

Algorithme 4 Regroupement des fenêtres se chevauchant en utilisant l'algorithme de RNN

Commencer une chaine C_1 avec une fenêtre aléatoire $f \in F$
Sauvegarder l'ensemble des fenêtres construites (F) dans la liste C_2
$dern \leftarrow 0; dernChev[0] \leftarrow 0; C_1[dern] \leftarrow f \in D; C_2 \leftarrow F \backslash f$
Tantque $C_2 \neq \emptyset$ **Faire**
  $(s, chev) \leftarrow plusGrandChevauchement(C_1[dern], C_2)$
  Si $chev > dernChev[dern]$ **Alors**
    $dern \leftarrow dern + 1; C_1[dern] \leftarrow s; C_2 \leftarrow C_2 \backslash \{s\}; dernChev[dern] \leftarrow sim$
  Sinon
    Si $dernChev[dern] > t_{chev}$ **Alors**
      $s \leftarrow regroupement(C_1[dern], C_1[dern-1])$
      $C_2 \leftarrow C_2 \cup \{s\}$
      $dern \leftarrow dern - 2$
    Sinon
      $dern \leftarrow -1$
    Fin Si
  Fin Si
  Si $dern < 0$ **Alors**
    Initialiser une nouvelle chaine avec une autre fenêtre aléatoire $f \in C_2$
    $dern \leftarrow dern + 1$
    $C_1[dern] \leftarrow f \in C_2; C_2 \leftarrow C_2 \backslash \{f\}$
  Fin Si
Fin Tantque

Les régions d'intérêt englobant les piétons sont déterminées en deux étapes. Dans la première, les fenêtres construites autour des POI sont regroupées selon

un taux de chevauchement t_{chev}. Vu que les taux d'occultation considérés est jusqu'à 60%, nous avons choisi d'attribuer à t_{chev} la valeur de 0.6. Les fenêtres se chevauchant ont été regroupées en adaptant le même algorithme de clustering agglomératif (RNN, voir section 3.2.1) utilisé pour créer le VV. Cette fois, l'algorithme RNN n'est pas utilisé pour grouper des descripteurs de POI, mais des fenêtres se chevauchant. Nous notons que le voisin le plus proche d'une fenêtre est celui qui maximise le pourcentage de recouvrement (ratio entre l'intersection et l'union). Les détails de la procédure de regroupement des fenêtres se chevauchant sont donnés dans l'algorithme 4.
Dans la deuxième étape, l'ensemble du corps du piéton est estimé de manière approximative en utilisant un ratio largeur/hauteur, estimé sur la base d'image d'apprentissage ($hauteur = 2.35 \times largeur$). Ensuite, les positions des fenêtres sont affinées par la recherche de la ligne qui contient plus de contours horizontaux marquant la transition entre le fond de l'image et le pied du piéton.

5.2.3 Validation des hypothèses

Après la construction de ROI, une étape de classification par SVM est utilisée afin de valider ou non la présence de piétons. Les résultats de mise en correspondance entre les POI et le VVH établis dans l'étape de détection sont réutilisés afin de caractériser localement les ROI. Ainsi, le processus d'extraction de caractéristiques se limite à calculer des descripteurs globaux afin de caractériser globalement les ROI. Enfin, l'ensemble de caractéristiques locales et globales extraites est évalué par SVM. Bien évidemment, les hypothèses classifiées en non piéton sont rejetées. Dans la figure 5.3 nous illustrons les résultats de détection de piétons obtenus après l'exécution de chaque étape de l'algorithme proposé.

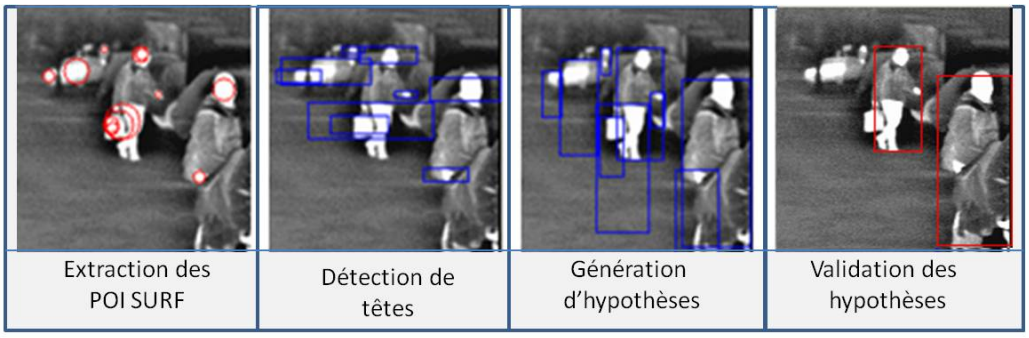


Figure 5.3. Illustration des résultats de détection obtenus après chaque étape de l'algorithme proposé

5.2.4 Suivi des piétons

Un algorithme de suivi, basé sur la mise en correspondance temporelle des descripteurs SURF, est enclenché pour les piétons détectés. Ce processus a été mis en œuvre non simplement pour éviter la lourde tâche de re-détection dans chaque image, mais aussi pour prendre le relais si le processus de détection échoue temporairement. De plus, le suivi permettra de pallier l'inconvénient majeur de l'algorithme de détection qui ne permet pas de détecter des piétons dans le cas d'occultation de tête.
L'algorithme de suivi que nous proposons est simple et rapide. Il n'utilise aucun filtre temporel et ne demande pas une étape de mise à jour, souvent gourmande en temps de calcul. Il est basé, tout simplement, sur la mise en correspondance temporelle entre les descripteurs SURF. Le processus de suivi se déroule en trois étapes :

1. **L'appariement de descripteurs :**
   Un piéton détecté dans l'image (g) est encadré par une fenêtre qui englobe un ensemble de descripteurs. Pour chacun d'eux, une recherche d'homologue est lancée dans l'image $(g+1)$. La recherche est effectuée dans une région d'intérêt dont la taille est proportionnelle à la taille de la fenêtre qui englobe le piéton dans l'image (g). Ensuite le descripteur le plus proche est déterminé en calculant la distance euclidienne entre descripteurs. Notons que deux seuils ont été utilisés afin d'assurer le bon fonctionnement du processus d'appariement. Un premier seuil a été utilisé afin d'évaluer la similarité entre les deux descripteurs. Quant au deuxième seuil, il a été utilisé afin de vérifier qu'aucun autre descripteurs de l'image $(g+1)$ ne permette de fournir une mesure de similarité proche. Les différents seuils utilisés ont été fixés après une série d'expérimentations.
2. **Génération de votes :**
   Chaque couple de points (k_g, k_{g+1}) appariés vote pour la nouvelle position du piéton avec un score basé sur leur mesure de similarité $S(k_g, k_{g+1})$. Un vote est représenté par le triplet (x, y, s). Le couple (x, y) désigne la nouvelle position du piéton dans l'image $(g+1)$. Le paramètre s, quant à lui, correspond à la différence d'échelle entre les deux objets. La figure 5.4 illustre le principe de génération d'une ROI à partir d'un couple apparié de POI.
3. **Génération de votes :**
   Chaque couple de POI apparié vote en 3D pour la nouvelle position du piéton

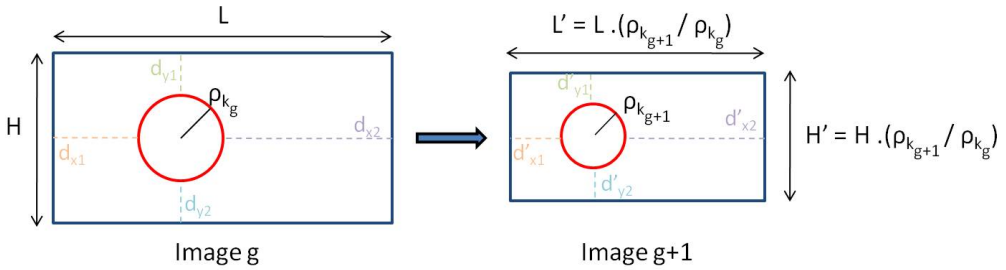


Figure 5.4. Génération d'une ROI suite à l'appariement du couple (k_g, k_{g+1}). La nouvelle fenêtre (dans l'image $g+1$) est positionnée de la même manière que la fenêtre initiale tout en respectant l'emplacement et l'échelle des POI appariés.

avec un score $S(k_g, k_{g+1})$. Ces votes sont ensuite interprétés par l'algorithme Mean Shift en 3D. L'avantage de cet algorithme est qu'il est très peu coûteux en temps de calcul (voir section 1.2.2.3). Il permet de trouver les coordonnées de la fenêtre englobante (position et échelle) dans une nouvelle image de manière itérative jusqu'à ce qu'il y ait convergence ou jusqu'à ce qu'un nombre maximum d'itérations soit atteint. Finalement, la nouvelle position du piéton est déterminée, puis validée par SVM.
La figure 5.5 illustre le principe de l'algorithme de suivi proposé.

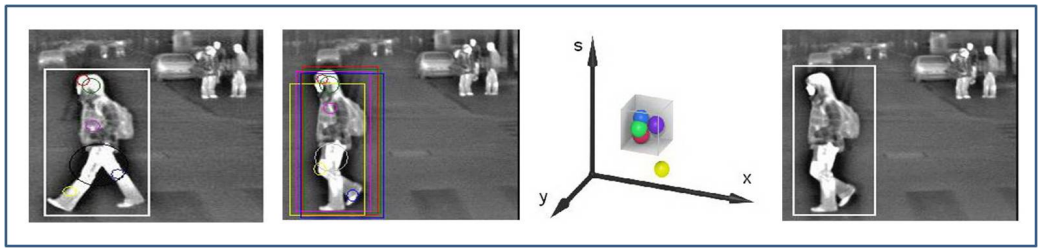


Figure 5.5. Illustration du principe de l'algorithme de suivi. La première image contient un piéton détecté et un ensemble de POI entourés par des cercles, dont les rayons correspondent à leurs valeurs d'échelles. Dans l'image qui suit, chaque région d'intérêt est construite après l'appariement temporel des descripteurs. Chaque couple de POI apparié vote pour la position et l'échelle du piéton dans l'image suivante. L'ensemble des votes est traité en 3D par l'algorithme Mean Shift qui fournit en sortie les coordonnées optimales de l'emplacement du piéton (dernière image)

5.2.5 Optimisation du temps de calcul

La diminution du temps de calcul, pour l'implémentation d'un système embarqué de détection de piétons, est un critère très important. La complexité de l'algorithme de détection est fortement dépendante du nombre de POI extraits

de chaque image. Ainsi, nous proposons d'effectuer un seuillage des valeurs hessiennes des POI afin de ne considérer que les régions très claires dans l'image. Le seuillage est naturellement fait en considérant une valeur seuil qui dépend de la valeur maximale d'Hessien pour chaque image. Ce choix est justifié par le fait que les zones caractéristiques de présence de piétons sont habituellement les plus claires. Ainsi, seulement les POIs vérifiant la contrainte ($H > H_{max} * t_{hes}$) seront retenus. Notons qu'il s'agit d'un seuillage adaptatif vu que H_{max} désigne la valeur maximale hessienne d'un POI extrait dans l'image en question. L'influence du coefficient t_{hes} sera étudiée dans la partie d'expérimentations.

5.3 Expérimentations et évaluations

Comme nous l'avons évoqué, la mise en place d'un système de détection de piétons embarqué à bord d'un véhicule est confrontée aux contraintes temps réel et aux problèmes liés principalement à la variabilité de l'apparence et de la forme des piétons. Dans cette section, nous étudions la robustesse et la précision du système de détection de piétons proposé. De plus, nous allons évaluer sa sensibilité au paramétrage initial.
Les résultats ont été obtenus sur deux séquences d'images de piétons appelées *Tetra1* et *Tetra2*. Le tableau 5.1 met en valeur la grande variabilité des résolutions des imagettes de piétons et donne des informations sur les taux d'occultation des piétons.

Table 5.1. Présentation des deux séquences utilisées pour les expérimentations

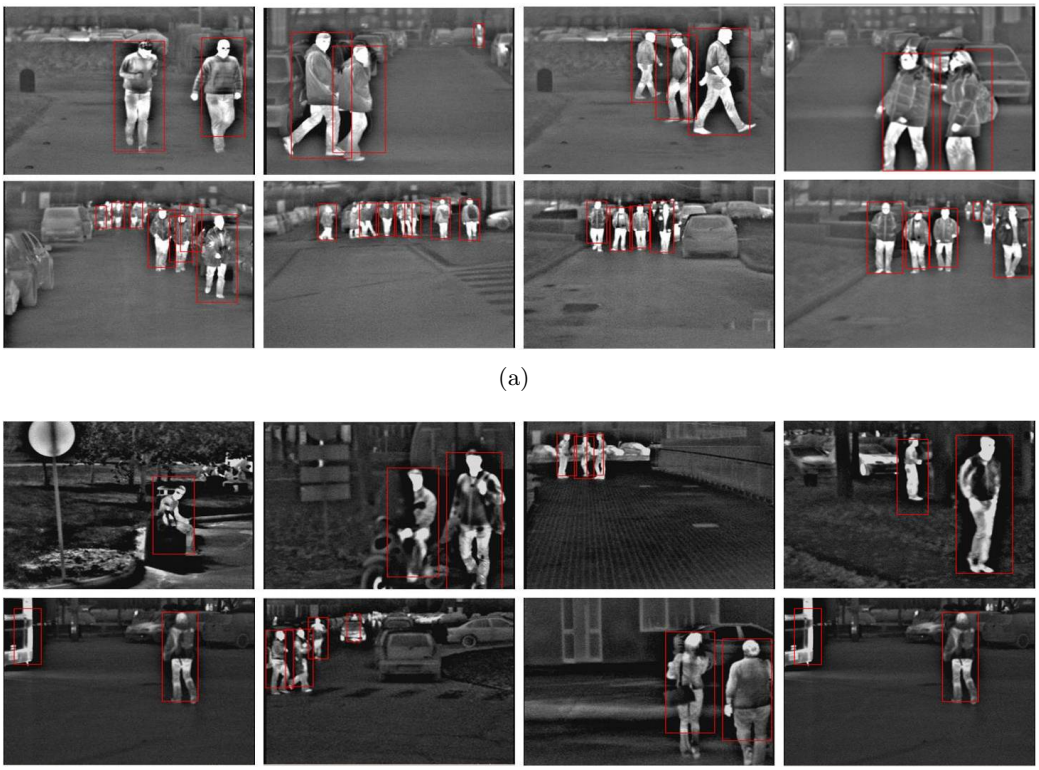
Séquence	nombre de piétons	plage des résolutions	résolution moyenne	écart-type des résolutions	Taux d'occultation
Tetra1	366	[133,24639]	6614.51	5525.17	13.66
Tetra2	454	[320,32640]	7926.12	5019.54	17.4

Dans ce qui suit, nous présentons les performances globales du système proposé de détection et de suivi.

5.3.1 Performances globales

La figure 5.6 présente des exemples de détection en présence de quelques occultations partielles. Il est clair que la majorité des piétons sont correctement détectés, indépendamment de leurs résolutions.

Afin d'évaluer les performances de notre système, nous réalisons une comparaison entre les résultats obtenus et les données expérimentales de référence (figure



(a)

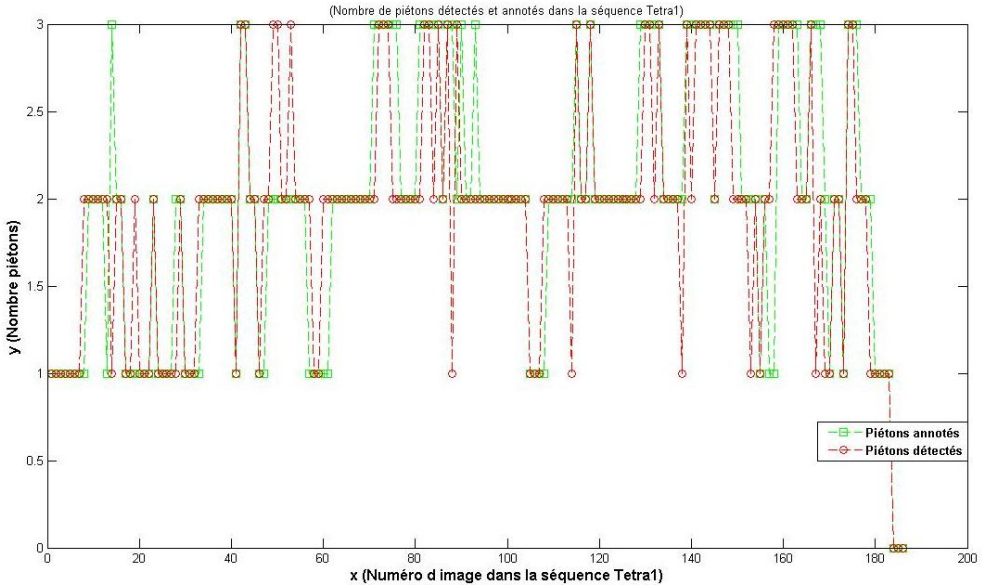
(b)

Figure 5.6. Quelques exemples de détections dans les deux séquences d'IR lointain *Tetra1* (fig.a) et Tetra2(fig.b). Toutes les images ont été traitées à leur résolution d'origine (320×240 pixels). Les résultats confirment la précision du système de détection même en présence d'occultations partielles.

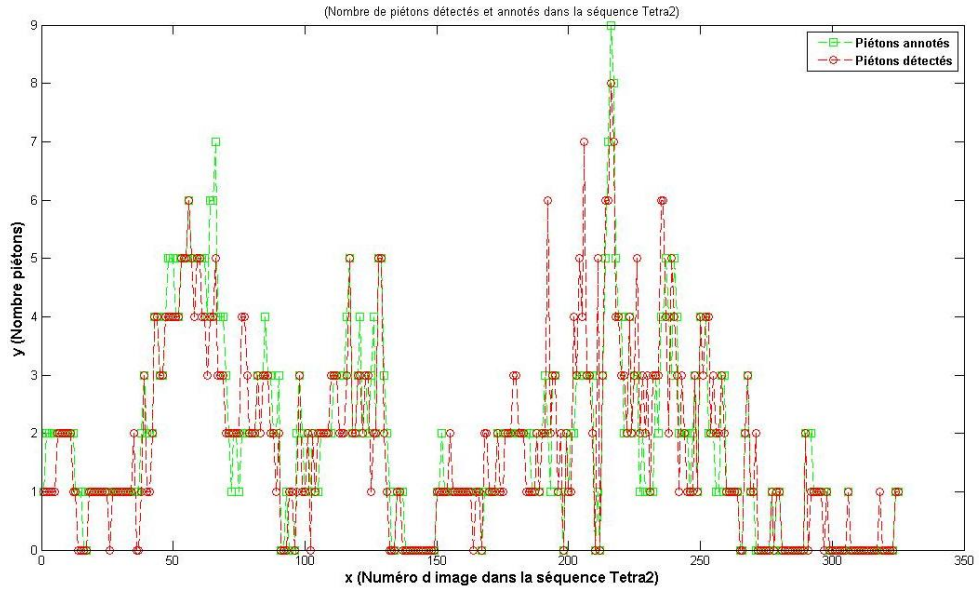
5.7). En effet, les courbes comparent, pour les deux séquences, le nombre de piétons réels et le nombre de piétons détectés. La figure montre que les courbes sont plus au moins superposées, ce qui indique que les résultats de détection sont significatifs.

Après une illustration graphique des résultats (5.7), nous récapitulons les résultats obtenus dans le tableau 5.2.

Le tableau 5.2 récapitule les performances de l'algorithme proposé et évalue l'importance du processus de suivi ainsi que la taille du vecteur descripteur. Les résultats présentés sont satisfaisants et montrent que l'algorithme proposé présente un bon compromis de F-mesure/temps de calcul. À l'opposé des résultats présentés dans les chapitres précédents, le meilleur taux de détection présenté est de 88.23%. Nous considérons que ces résultats sont satisfaisants car la tâche est plus difficile puisqu'elle consiste à localiser, puis à identifier les piétons dans les images. Dans les chapitres précédents, tous les résultats présentés ont été obtenus en analysant des fenêtres englobantes annotées manuellement.



(a)



(b)

Figure 5.7. Comparaison entre les résultats obtenus et les données expérimentales de référence. Les courbes comparent, pour les deux séquences, le nombre de piétons réels et le nombre de piétons correctement détectés.

Les comparaisons résumées dans le tableau ne font pas apparaître de différences significatives quant aux taux F-score obtenus en utilisant les descripteurs SURF-64 et SURF-128. En revanche, nous constatons une réduction significative du temps

Table 5.2. Résultats de détection et de suivi en IR lointain

Séquence	Descripteur	Suivi	F-score (%)	Temps de calcul (image / ms)
Tetra1	SURF-64	Sans	83,88	92
		Avec	88,07	91.2
	SURF-128	Sans	84,71	125
		Avec	88,23	125.3
Tetra2	SURF-64	Sans	76,08	92
		Avec	81,08	131
	SURF-128	Sans	76,53	104
		Avec	81,56	104.8

de calcul obtenu en utilisant SURF-64. Ceci est cohérent avec les observations faites dans le chapitre 3 (voir section 3.3.3.4). En ce qui concerne le suivi temporel, les résultats montrent l'impact positif de ce processus. En effet, il a permis non seulement d'améliorer les résultats de détection mais aussi de maintenir un temps de calcul acceptable. Il est fort de mentionner que le système proposé, en termes de temps de traitement, est rapide puisqu'il permet de traiter environ 10 images par seconde. Cela révèle l'importance de la technique de filtrage des POI SURF proposée dans la section 5.2.5. Dans le reste du chapitre, nous étudions l'influence des différents paramètres sur les performances globales du système.

5.3.2 Influence du paramétrage

Dans cette section, nous étudions la sensibilité du système de détection et de suivi au paramétrage initial. Rappelons que les différents paramètres du système sont le seuil de sélection des POI SURF (t_{hes}), le seuil de chevauchement (t_{chev}) et la profondeur du VVH.

5.3.2.1 Seuillage des POI SURF

Nous étudions dans cette section, la sensibilité de l'algorithme proposé par rapport au nombre de POI SURF extraits des images. Rappelons que pour des considérations du temps de calcul, nous avons proposé de seuiller les valeurs hessiennes des POI. Ainsi seulement les POIs très contrastés par rapport au fond d'images sont retenus. Dans la figure 5.8, nous étudions l'influence de la valeur du seuil t_{hes} utilisé afin de sélectionner seulement les POI ayant une valeur hessienne H vérifiant ($H > H_{max} * t_{hes}$).

Les expérimentations schématisées dans la figure 5.8 montrent que le temps

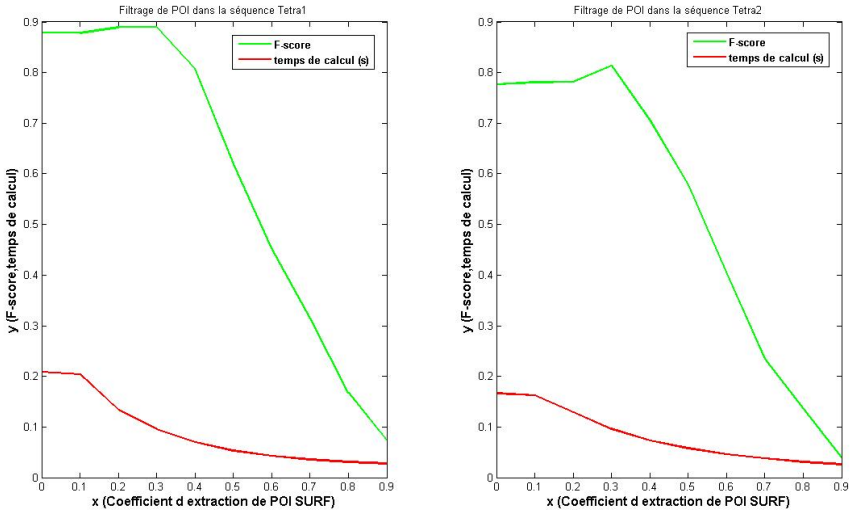


Figure 5.8. Influence du coefficient de seuillage utilisé

de calcul est fortement lié au nombre de POI SURF extraits de chaque image. En effet, la courbe en rouge montre que le seuillage des POIs permet de réduire significativement le temps de calcul. En revanche, la courbe verte montre qu'à partir d'une certaine valeur ($t_{hes} = 0.3$), les résultats de détection se dégradent. Cela est dû principalement au fait que plusieurs POI de piétons ont été éliminés. Ainsi, nous pouvons conclure que ce paramètre, bien qu'il ait un avantage évident sur le temps de calcul, nécessite un ajustement très précis. Il est important de noter que les résultats présentés précédemment ont été obtenus avec ($t_{hes} = 0.3$). Cette valeur, estimée sur l'ensemble de validation, a été fixée à un compromis entre le taux de F-mesure et le temps de calcul.

5.3.2.2 Seuil de chevauchement

Le seuil de chevauchement est également un paramètre déterminant quant aux performances de détection. Ce seuil (t_{chev}) a été utilisé principalement afin de regrouper les fenêtres se chevauchant. Le regroupement s'est déroulé en trois temps afin de regrouper les hypothèses qui ont été générées suite aux processus de :

- détection des têtes,
- détection du corps entier des piétons,
- comparaison et fusion entre les hypothèses provenant des étapes de détection

et de suivi.

Il est important de mentionner que nous considérons un taux maximal d'occultation de 60%. En d'autres termes, si le taux de recouvrement (ratio entre l'intersection et l'union) dépasse cette valeur, seulement l'objet résultant de la fusion est pris en considération. Cette contrainte a été vérifiée durant l'étape d'annotation. C'est pour cette raison que nous avons fixé la valeur de (t_{chev}) à 0.6. Dans la figure 5.9, nous examinons l'influence de ce seuil.

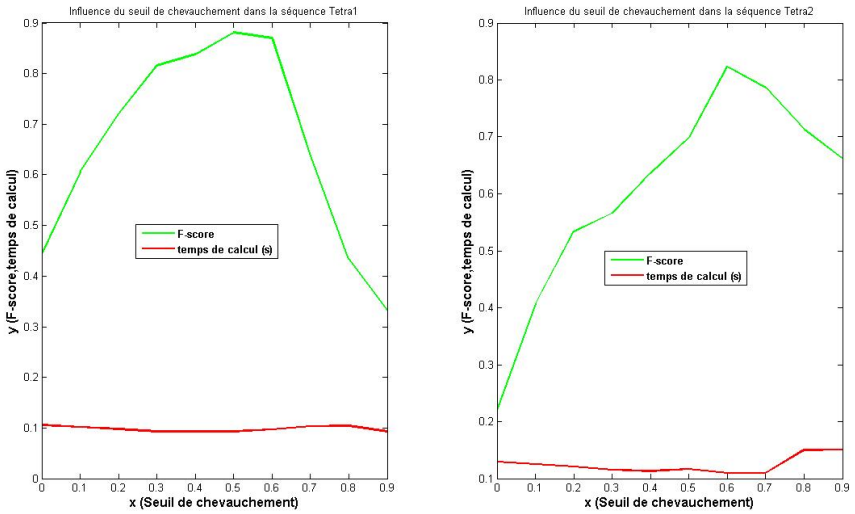


Figure 5.9. Influence de la valeur du seuil de chevauchement

Les allures des courbes de F-score (figure 5.9) montrent que les performances du système de détection proposé sont fortement dépendantes du seuil de chevauchement fixé au préalable. Toutefois, il est intéressant de constater que la valeur optimale de ce seuil est très proche du taux maximal d'occultation considéré. Au dessous de cette valeur, les résultats présentés ne sont pas satisfaisants car le système a tendance à produire des faux positifs. Au dessus de cette valeur, les résultats se dégradent de manière significative. Cela est certainement dû au regroupement des fenêtres encadrant des objets distincts. En ce qui concerne les temps de calcul, la figure 5.9 montre que les données temporelles sont inversement proportionelles aux taux de F-score. Cette observation est particulièrement marquée dans les résultats issus du traitement de la deuxième séquence *Tetra2*. Ceci est expliqué par le fait que le nombre des piétons occultés dans la séquence *Tetra2* est plus important que pour la séquence de Tetra1.

Finalement, nous pouvons conclure que d'une part, l'impact de ce seuil est extrêmement important. D'autre part, l'ajustement de sa valeur par le seuil d'occultation considéré, permet d'assurer le bon fonctionnement du système de détection.

5.3.2.3 Profondeur du Vocabulaire Visuel

Dans la section 3.3.3.5, nous avons montré que la profondeur du VV a un impact considérable sur les performances de reconnaissance de piétons. Le système de détection proposé ici, inclut une étape de génération de ROI et de reconnaissance. Etant donné que ces deux étapes sont basées sur l'utilisation du VVH, l'examen de l'impact de la structure du VVH est indispensable.

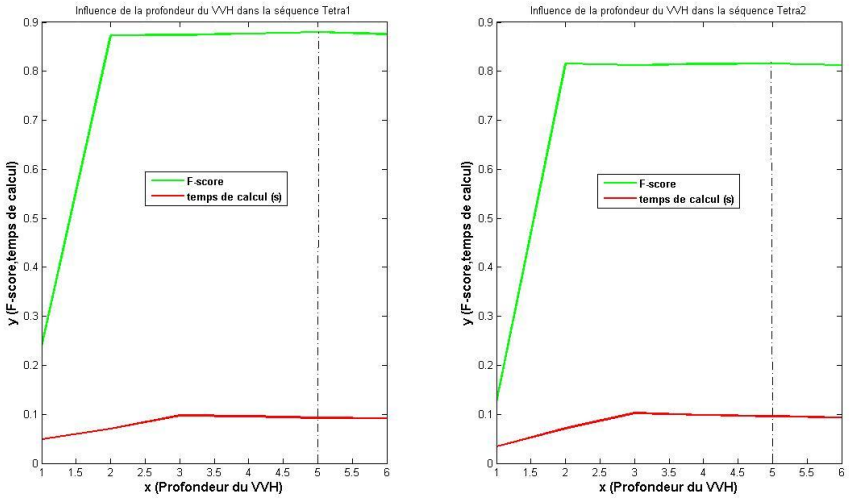


Figure 5.10. Influence de la profondeur du VVH

Nous présentons dans la figure 5.10 l'évolution des scores de détection en fonction de la profondeur du VVH. La profondeur 5 fournit le meilleur compromis entre taux de F-score et temps de calcul. Cette observation est cohérente avec les résultats présentés dans la section 3.3.3.5. Ainsi, cela confirme davantage la pertinence de la mesure d'évaluation proposée lors de la construction du VVH (voir section 3.2). Semblablement à ce qui a été observé dans la section 3.3.3.5, la courbe montre que des améliorations significatives ont été apportées grâce à la structure hiérarchique du VV. En ce qui concerne les temps de calcul, les courbes montrent qu'à partir de la profondeur 3, se révèle l'impact de la structure hiérarchique sur l'accélération des temps de mise en correspondance. Les résultats de F-score en-

registrés jusqu'à la prondeur 3, montrent que l'augmentation du nombre de la profondeur du VVH ne commence à donner naissance à des clusters significatifs qu'à partir de cette profondeur.

5.3.3 Bilan

Dans cette section, nous avons évalué les performances globales du système de détection proposé. Les résultats obtenus sur des images routières de piétons en milieu urbain ont permis de valider le système avec un taux de détection satisfaisant. En effet, le système atteint un taux de F-score moyen de 84% et permet de traiter environ 10 images par seconde. Conformément aux résultats présentés dans le chapitre 3, les expérimentations ont montré également que la structure hiérarchique du Vocabulaire Visuel a permis non seulement d'accélérer les temps de calcul mais aussi d'améliorer les performances de détection. De plus, nous avons présenté une étude permettant en particulier d'évaluer l'impact des différents paramètres du système sur ses performances.
Enfin, nous ne pouvons pas conclure sans mentionner que plusieurs travaux d'une équipe italienne de renom (Laboratoire VisLab) [BBFV06] ont été menés sur les mêmes bases d'images utilisées pour les expérimentations. Nous regrettons de ne pas être en mesure de comparer les résultats vu que les images mises à notre disposition n'étaient pas pourvues d'annotations. Toutefois, le fait que les autres travaux se sont basés sur une combinaison d'une paire de systèmes stéréos VIS et IR, permet de révéler la difficulté des images traitées.

5.4 Conclusion

Dans ce chapitre, nous avons présenté un système original de détection de piétons dans les images d'infrarouge lointain. Le système procède à trois étapes : l'apprentissage, la détection et le suivi, qui se basent toutes sur l'extraction et l'appariement des points d'intérêt SURF à partir des régions claires dans l'image. Les expérimentations montrent que le système proposé produit des résultats précis et robustes face aux problèmes de changements d'échelle et d'occultations partielles des piétons.
Ce chapitre clôt donc le raisonnement scientifique formulé dans ce mémoire articulé autour des trois mouvements principaux : la problématique de la reconnaissance d'objets, l'intérêt de la fusion puis l'application à la détection de piétons.

Conclusion et perspectives

Bilan

Cette thèse s'inscrit dans le contexte de la vision embarquée pour la détection et la reconnaissance d'obstacles routiers, en vue d'application d'assistance à la conduite automobile. Composés de systèmes visant à assister le conducteur sur différentes dimensions de la conduite, les applications d'assistance à la conduite présentent un enjeu majeur du point de vue sécurité routière.
L'intégration d'un système de détection d'obstacles routiers nécessite avant tout la maîtrise des technologies de perception de l'environnement du véhicule. La complexité de la tâche de perception provient d'une part de la grande diversité des scénarios routiers, d'autre part, de la large variabilité des formes et des apparences des obstacles routiers. Notre choix s'est porté sur un système en monovision afin de pouvoir implémenter une technique à la fois rapide et peu onéreuse. Afin de surmonter les conditions de visibilité réduite, surtout la nuit, nous avons choisi d'utiliser une caméra Infrarouge (LWIR) embarquée. En revanche, d'autres difficultés liées à l'interprétation systématiques des images routières en milieu urbain, restent à surmonter. Identifier, analyser et résoudre les difficultés plus particulièrement liées aux applications de détection et de reconnaissance d'obstacles routiers sont les objectifs premiers de cette thèse.
Après une étude bibliographique, rapportée dans le premier chapitre, nous avons constaté que la problématique de détection et de suivi d'obstacles routiers dans des scènes dynamiques, ne peut être résolue convenablement sans recourir aux techniques de reconnaissance de catégories d'objets dans les images. Pour répondre à cette problématique, nous avons axé notre réflexion autour de trois axes : la représentation, la classification et la fusion d'informations.
Le deuxième chapitre de la thèse expose un état de l'art qui s'articule autour de ces trois axes. Pour chacun, nous avons examiné les méthodes les plus couramment employées dans notre contexte applicatif en écartant celles moins adaptées à nos

besoins. Après avoir présenté la problématique (chapitre 1) et l'état de l'art (chapitre 2), nous avons exposé les choix opérés et les méthodes mises en œuvre afin de répondre à la problématique de détection et de reconnaissance des obstacles routiers.

Une partie de la réponse à la problématique a pu être apportée par un choix pertinent de la représentation. Ainsi, nos travaux ont porté dans un premier temps sur l'étude des méthodes de caractérisation adéquates. À cet égard, nous avons proposé un modèle de représentation basé non seulement sur une caractérisation locale mais aussi sur une caractérisation globale permettant ainsi de résoudre le problème de variabilité des formes et des apparences des obstacles routiers. La caractérisation globale résulte de l'extraction de caractéristiques à partir d'une imagette permettant de décrire les formes et les textures globales des obstacles. La caractérisation locale, quant à elle, consiste à extraire une signature de l'apparence locale d'un obstacle par la mise en correspondance de descripteurs locaux (SURF) avec un Vocabulaire Visuel Hiérarchique. La structure hiérarchique a été conçue non seulement pour accélérer le processus de mise en correspondance mais aussi pour gérer différents niveaux de similarité, ce qui autorise une grande souplesse de représentation. Dans un deuxième temps, nous avons proposé une méthode permettant de combiner le modèle d'apparence avec une technique de classification afin de catégoriser précisément les obstacles routiers.

Différentes expérimentations ont été menées afin d'aboutir au choix des différents composants du système de reconnaissance proposé. En outre, nous avons confronté notre système avec d'autres fondés sur : des descripteurs (SIFT), des fonctions noyaux spécifiques pour la mise en correspondance de descripteurs locaux (LMK) et d'autres méthodes de caractérisation telles que les ondelettes de Haar et de Gabor. Les résultats expérimentaux obtenus montrent l'intérêt de notre système de reconnaissance qui présente les meilleurs résultats à partir des images infrarouges, notamment pour la classe Piéton. En revanche, les résultats de reconnaissance de véhicules ont été moins bons. Cela a été justifié par le fait que les véhicules, notamment stationnés, ne présentent pas un contraste suffisant pour être repérés dans les images infrarouges.

Le deuxième axe que nous avons souligné est l'apport de la fusion multimodale. Cet axe a été exploré non seulement pour améliorer les performances globales du système, mais également pour mettre en jeu la complémentarité des caractéristiques locales et globales ainsi que les deux modalités visible et infrarouge. Les deux catégories de caractéristiques (locales et globales) extraites à partir des deux

modalités (visible et infrarouge) ont été amenées à être vues comme des sources d'informations à combiner. Pour réduire la complexité du système et dans le but de respecter la contrainte temps réel, une stratégie de classification à deux niveaux de décision a été proposée. Cette stratégie est basée sur la théorie des fonctions de croyance (théorie de Dempster-Shafer) et permet d'accélérer grandement le temps de prise de décision. Bien que la fusion de capteurs ait pu améliorer les résultats de reconnaissance, nous avons constaté que l'apport est particulièrement faible pour la classe Piéton. Quoi qu'il en soit, les résultats obtenus sont satisfaisants, mais pas assez robustes pour assurer l'implantation d'un système de détection générique. Cela nous a conduit à envisager l'implantation d'un seul capteur infrarouge pour la mise en œuvre d'un système embarqué de détection de piéton.
Dans le dernier chapitre, nous avons mis à profit les résultats d'expérimentations et nous avons intégré les éléments développés dans un système de détection et de suivi de piéton en infrarouge-lointain. Ceci a été réalisé en injectant dans un premier temps, une information spatiale implicite au sein du Vocabulaire Visuel Hiérarchique et en utilisant, dans un deuxième temps, un classifieur SVM pour valider les hypothèses de détection et de suivi. Ce système a été validé aux travers différentes expérimentations sur des images et des séquences routières dans un milieu urbain. Les expérimentations montrent que le système proposé est rapide et produit des résultats satisfaisants face aux problèmes de changements d'échelle et d'occultations partielles.

Limites des méthodes proposées

Les travaux présentés dans cette thèse explorent un ensemble de techniques visant à détecter et à reconnaître des obstacles dans des scènes routières. Pour certaines d'entre elles, nos travaux peuvent être améliorés.
Le premier point à améliorer consiste à étudier les nouvelles solutions d'indexation rapides proposées dans le domaine de recherche d'information afin de positionner l'efficacité et la rapidité du processus de caractérisation locale. Le Vocabulaire Visuel Hiérarchique proposé ne code pas des relations spatiales entre les noeuds. Ainsi, une piste prometteuse à explorer consisterait à enrichir la représentation des apparences locales dans le Vocabulaire Visuel en intégrant des informations de localisation spatiale. Cela permettra d'intégrer d'imposer des contraintes spatiales qui pourraient améliorer le processus de mise en correspondance. En ce qui concerne la classification par SVM, il faudrait proposer des formulations de noyaux

plus adaptées à cette nouvelle structure.
Le deuxième point à améliorer concerne l'algorithme de détection de piétons proposé. Cet algorithme présente l'inconvénient de ne pas parvenir à détecter un piéton dont la tête est cachée. Ce problème n'a pas été observé dans l'ensemble des images utilisées. Bien que, le processus de suivi temporel permettait de retrouver la nouvelle position du piéton même en cas d'éventuelle occultation de tête, il faudrait étudier des situations particulières comme, par exemple, le maintien d'une parapluie. Une solution qui pourrait être envisagée serait d'inclure dans le Vocabulaire Visuel des points d'intérêt extraits à partir d'objets particuliers qui pourraient masquer les têtes des piétons.
Enfin, il faudrait faire d'autres expérimentations afin d'évaluer l'impact des croisements des piétons lors du suivi de plusieurs personnes.

Perspectives

Après avoir déterminé et analysé les limites des méthodes proposées dans ce travail de thèse, nous présentons les perspectives envisageables de nos travaux de recherche. Les pistes à explorer et les applications sont nombreuses :
- Extension du système de reconnaissance pour qu'il englobe d'autres objets routiers comme les panneaux routiers, les animaux, ...
- Calibrage automatique en temps réel d'une caméra visible et d'une autre infrarouge afin d'envisager un processus de fusion en amont de la phase de reconnaissance.
- Enfin, une perspective à long terme est de proposer des fonctions avancées d'aides à la conduite, comme les régulateurs de vitesse adaptatif (ACC) en remplaçant les radars utilisés aujourd'hui dans les nouvelles voitures, par des caméras embarquées.

Pour conclure, il semble que le domaine de détection d'obstacles routiers par des caméras reste toujours actif malgré le progrès technique et la recherche avancée des systèmes de communication entre véhicules et avec infrastructure. La recherche dans le domaine des caméras embarquées est encore d'actualité et devrait le rester avant d'éteindre la fiabilité exigée par l'industrie.

Bibliographie

[AAB+02] L. Andreone, P.C Antonello, M. Bertozzi, A. Broggi, A. Fascioli, and D. Ranzato. Vehicle detection and localization in infra-red images. *IEEE International Conference on Intelligent Transportation Systems*, pages 141–146, 2002.

[AAR04] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 :1475–1490, 2004.

[App91] A. Appriou. Probabilités et incertitude en fusion de données multi-senseurs. *Revue scientifique et technique de la défense*, 11 :27–40, 1991.

[ARB08] A. Apatean, A. Rogozan, and A. Bensrhair. Kernel and feature selection for visible and ir based obstacle recognition. *IEEE Conference on Intelligent Transportation Systems*, pages 1130–1135, 2008.

[ARB09a] A. Apatean, A. Rogozan, and A. Bensrhair. Obstacle recognition using multiple kernel in visible and infrared images. *IEEE Intelligent Vehicles Symposium*, pages 370–375, 2009.

[ARB09b] A. Apatean, A. Rogozan, and A. Bensrhair. Svm-based obstacle classification in visible and infrared images. 2009.

[ASDT+07] D.F. Alonso, I.P.and Llorca, L.M. Sotelo, M.A.and Bergasa, P.R. Del Toro, J. Nuevo, M. Ocana, and M.A.G. Garrido. Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2) :292–307, 2007.

[Aya07] S. Ayache. *Indexation de documents vidéos par concepts par fusion de caractéristiques audio, image et texte.* PhD thesis, Institut National Polytechnique de grenoble, 2007.

[BAC06] T. Burger, O. Aran, and A Caplier. Modeling hesitation and conflict : A belief-based approach for multi-class problems. *Proceedings of the 5th International Conference on Machine Learning and Applications*, pages 95–100, 2006.

[BBDR+07] M. Bertozzi, A. Broggi, M. Del Rose, M. Felisa, A. Rakotomamonjy, and F. Suard. A pedestrian detector using histograms of oriented gradients and a support vector machine classifier. *IEEE Intelligent Transportation Systems Conference*, pages 143–148, 2007.

[BBF+04] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf, and M.M. Meinecke. Pedestrian detection for driver assistance using multiresolution infrared vision. *IEEE transactions on vehicular technology*, 53(6) :1666–1678, 2004.

[BBFL02] M. Bertozzi, A. Broggi, A. Fascioli, and P. Lombardi. Vision-based pedestrian detection : will ants help ? *IEEE Intelligent Vehicles Symposium*, pages 1–7, 2002.

[BBFN00] M. Bertozzi, A. Broggi, A. Fascioli, and S. Nichele. Stereo vision-based vehicle detection. *IEEE Intelligent Vehicles Symposium*, pages 39–44, 2000.

[BBFV06] M. Bertozzi, A. Broggi, M. Felisa, and G. Vezzoni. Low-level pedestrian detection by means of visible and far. *IEEE Intelligent Vehicles Symposium*, pages 231–236, 2006.

[BBG+07] M. Bertozzi, A. Broggi, C.H. Gomez, R.I. Fedriga, G. Vezzoni, and M. Del Rose. Pedestrian detection in far infrared images based on the use of probabilistic templates. *IEEE Intelligent Vehicles Symposium*, pages 327–332, 2007.

[BC09] T. Burger and A. Caplier. A generalization of the pignistic transform for partial bet. *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 252–263, 2009.

[Bea78] P.R. Beaudet. Rotationally invariant image operators. *International Joint Conference on Pattern Recognition*, pages 579–583, 1978.

[BHD00] M. Betke, E. Haritaoglu, and L. S. Davis. Real-time multiple vehicle detection and tracking from a moving vehicle. *Machine vision and applications*, 12(2) :69–83, 2000.

[Blo03] I. Bloch. Fusion d'informations en traitement du signal et des images. *Hermes Science Publications*, 2003.

[BLRB10] B. Besbes, B. Labbe, A. Rogozan, and A. Bensrhair. Svm-based fast pedestrian recognition using a hierarchical codebook of local features. *IEEE Workshop on Machine Learning for Signal Processing*, pages 226–231, 2010.

[BLS09] B. Besbes, C. Lecomte, and P. Subirats. Nouvel algorithme de détection des lignes de marquage au sol. *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*, 2009.

[BMS09] T. Bdiri, F. Moutarde, and B. Steux. Visual object categorization with new keypointbased adaboost features. *IEEE Intelligent Vehicles Symposium*, pages 393–398, 2009.

[BP07] N. Bauer, J.and Sunderhauf and P. Protzel. Comparing several implementations of two recently published feature detectors. *International Conference on Intelligent and Autonomous Systems*, 2007.

[BRB10] B. Besbes, A. Rogozan, and A. Bensrhair. Pedestrian recognition based on hierarchical codebook of surf features in visible and infrared images. *IEEE Intelligent Vehicles Symposium*, pages 156–161, 2010.

[BTG06] H. Bay, T. Tuytelaars, and L.V. Gool. Surf : Speeded up robust features. *European Conference on Computer Vision*, pages 404–417, 2006.

[Can86] F.J. Canny. A computational approach to edge detection. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 8(6) :679–698, 1986.

[Can07] S. Canu. Machines à noyaux pour l'apprentissage statistique. *Techniques de l'ingénieur*, TE5255, 2007.

[CDF+04] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorizationwith bags of keypoints. *ECCV workshop on Statistical Learning in ComputerVision*, pages 59–74, 2004.

[Com03] D. Comaniciu. Non parametric information fusion for motion estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–66, 2003.

[CRM00] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, pages 142–149, 2000.

[CSY08] Y. Cui, L. Sun, and S. Yang. Pedestrian detection using improved histogram of oriented gradients. *International Conference on Visual Information Engineering*, pages 388–392, 2008.

[CTB05] I. Cabani, G. Toulminet, and A. Bensrhair. Color-based detection of vehicle lights. *IEEE Intelligent Vehicles Symposium*, pages 278–283, 2005.

[CZQ05] H. Cheng, N. Zheng, and J. Qin. Pedestrian detection using sparse gabor filter and support vector machine. *IEEE Intelligent Vehicles Symposium*, pages 583–587, 2005.

[DE84] W.H. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1) :7–24, 1984.

[Dem67] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2) :325–339, 1967.

[Den95] T. Denoeux. A k-nearest neighbour classification rule based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(1) :804–813, 1995.

[Den08] T. Denoeux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial Intelligence*, 172(2-3) :234–264, 2008.

[DGL07] J Dong, J Ge, and Y Luo. Nighttime pedestrian detection with near infrared using cascaded classifiers. *IEEE International Conference on Image Processing*, pages 185–188, 2007.

[Dis10] A. Discant. *Contributions à la fusion des informations. Application à la reconaissance des obstacle dans les images visible et infrarouge.* PhD thesis, l'Institut National des Sciences Appliquées de Rouen, 2010.

[DL97] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1 :131–156, 1997.

[DP92] D. Dubois and H. Prade. On the combination of evidence in various mathematical frameworks. *Reliability Data Collection and Analysis*, pages 213–241, 1992.

[DPKA04] C. Demonceaux, A. Potelle, and D. Kachi-Akkouche. Obstacle detection in a road scene based on motion analysis. *IEEE Transactions on Vehicular Technology*, 53(6) :1649–1656, 2004.

[DPS01] D. Dubois, H. Prade, and P. Smets. New semantics for quantitative possibility theory. *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 410–421, 2001.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[ELW03] H. Elzein, S. Lakshmanan, and P. Watta. A motion and shape-based pedestrian detection algorithm. *IEEE Intelligent Vehicles Symposium*, pages 500–504, 2003.

[ETLF11] W. Elloumi, S. Treuillet, R. Leconge, and A. Fonte. Performance evaluation of point matching methods in video sequences with abrupt motions. *International Conference on Computer Vision Theory*, pages 427–430, 2011.

[FC08] F. Fayad and V. Cherfaoui. Tracking objects using a laser scanner in driving situation based on modeling target shape. *IEEE Intelligent Vehicles Symposium*, pages 44–49, 2008.

[FH75] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1) :32–40, 1975.

[FLCS05] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. *International Conference on Computer Visio*, pages 1363–1370, 2005.

[FM09] A. Fiche and A. Marti. bayesienne et fonctions de croyance continues pour la classification. *Rencontres Francophones sur la Logique Floue et ses Applications*, 2009.

[FS96] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *International Conference on Machine Learning*, pages 148–156, 1996.

[Fuk90] K. Fukunaga. *K. Fukunaga.* Academic Press Professional, 1990.

[Gav00] D. Gavrila. Pedestrian detection from a moving vehicle. *Proceedings of European Conference on Computer Vision*, pages 37–49, 2000.

[GD07] K. Grauman and T. Darrell. The pyramid match kernel : Efficient learning with sets of features. *The Journal of Machine Learning Research*, 8 :725–760, 2007.

[GE03] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[GGB06] M. Grabner, H. Grabner, and H. Bischof. Fast approximated sift. *In proceedings of Asian Conference on Computer Vision*, pages 918–927, 2006.

[GGM04] D. M. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection : The protector system. *IEEE Intelligent Vehicles Symposium*, pages 13–18, 2004.

[GM07] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision (est : February-March*, 73 :41–59, 2007.

[GSS93] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F on Radar and Signal Processing*, 140, pages = 107-113, number = 2,, 1993.

[Hal98] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[HK09] A. Haselhoff and A. Kummert. A vehicle detection system based on haar and triangle features. *Intelligent Vehicles Symposium*, pages 261–266, 2009.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

[HT98] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *Annals of statistics*, 26(2) :451–471, 1998.

[JA09] K. Jungling and M. Arens. Feature based person detection beyond the visible spectrum. *Computer Vision and Pattern Recognition Workshop*, pages 30–37, 2009.

[Joa98] T. Joachims. Text categorization with support vector machines : Learning with many relevant features. *European Conference on Machine Learning*, pages 137–142, 1998.

[JVJS03] M. Jones, P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IEEE International Conference on Computer Vision*, pages 734–741, 2003.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82(82 (Series D)) :35–45, 1960.

[Kle08] J. Klein. *Suivi robuste d'objets dans les séquences d'images par fusion de sources, application au suivi de véhicules dans les scènes routières.* PhD thesis, Laboratoire LITIS - Université de Rouen, 2008.

[Kra08] S. Kramm. *Production de cartes éparses de profondeur avec un système de stéréovision embarqué non-aligné.* PhD thesis, Université de Rouen, 2008.

[KS04] Y. Ke and R. Sukthankar. Pca-sift : A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition*, pages 506–513, 2004.

[LBH08] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows : Object localization by efficient subwindow search. *Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[LCL09] L. Leyrit, T. Chateau, and J.T. Lapresté. Descripteurs pour la reconnaissance de piétons. In *Congrès des jeunes chercheurs en vision par ordinateur*, 2009.

[Lei08] A. Schiele B. Leibe, B. Ettlin. Learning semantic object parts for object categorization. *Image and Vision Computing*, 26(1) :15–26, 2008.

[LLS04] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. *ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[LM02] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. *International Conference on Image Processing*, pages 900–903, 2002.

[LM09] H. Laanaya and A. Martin. Fusion multi-vues à partir de fonctions de croyance pour la classification d'objets. In *Extraction et Gestion des Connaissances*, 2009.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision*, pages 1150–1157, 1999.

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[LSS05] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 878–885, 2005.

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1 :281–297, 1967.

[Mar05] A. Martin. Fusion de classifieurs pour la classification d'images sonar. *Revue Nationale des Technologies de l'Information*, pages 259–268, 2005.

[MCB06] D. Martins, M.R. Cesar, and J. Barrera. W-operator window design by minimization of mean conditional entropy. *Pattern analysis and applications*, 9(2-3) :139–153, 2006.

[Mer06] D. Mercier. *Fusion d'informations pour la reconnaissance automatique d'adresses postales dans le cadre de la théorie des fonctions de croyance*. PhD thesis, Université de technologie de compiègne, 2006.

[MLS06] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. *IEEE Conference on Computer Vision and Pattern Recognition*, 1 :26–36, 2006.

[Mor77] H.P. Moravec. Towards automatic visual obstacle avoidance. *International Joint Conference on Artificial Intelligence*, pages 584–584, 1977.

[MQD08] D. Mercier, B. Quost, and T. Denoeux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2) :246–258, 2008.

[MR04] M. Meis, U. Oberlander and W. Ritter. Reinforcing the reliability of pedestrian detection in far-infrared sensing. *Intelligent Vehicles Symposium*, pages 779–783, 2004.

[MS04] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1) :63–86, 2004.

[MS06] M. Marszalek and C. Schmid. Spatial weighting for bag-of-features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 :2118–2125, 2006.

[NCHP08] P. Negri, X. Clady, S.M. Hanif, and L. Prevost. A cascade of boosted generative and discriminative classifiers for vehicle detection. *EURASIP Journal on Advances in Signal Processing*, 136 :1–12, 2008.

[OPS+97] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 193–199, 1997.

[PLR+06] M. Perrollaz, R. Labayrade, C. Royere, N. Hautiere, and D. ; Aubert. Long range obstacle detection using laser scanner and stereovision. *IEEE Intelligent Vehicles Symposium*, pages 182–187, 2006.

[Por09] A. Porebski. *Sélection d'attributs de texture couleur pour la classification d'images. Application à l'identification de défauts sur les décors verriers imprimés par sérigraphie.* PhD thesis, Université Lille 1 - Sciences et technologies, 2009.

[PSZ08] S. Paisitkriangkrai, C. Shen, and J. Zhang. Fast pedestrian detection using a cascade of boosted covariance features. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8) :1140–1151, 2008.

[PTP98] C. Papageorgiou, Evgeniou ; T., and T. Poggio. A trainable pedestrian detection system. *IEEE lntelligent Vehicles Symposium*, pages 241–246, 1998.

[SAM+09] B. Schiele, M. Andriluka, N. Majer, S. Roth, and C. Wojek. Visual people detection - different models, comparison and discussion. *IEEE International Conference on Robotics and Automation*, 2009.

[SBM02] Z. Sun, G. Bebis, and R. Miller. On-road vehicle detection using gabor filters and support vector machines. *International Conference on Digital Signal Processing*, pages 1019–1022, 2002.

[SBM05] Zehang Sun, George Bebis, and Ronald Miller. On-road vehicle detection using evolutionary gabor filter optimization. *IEEE Transactions on Intelligent Transportation Systems*, 6 :125–137, 2005.

[SBM06] Z. Sun, G. Bebis, and R. Miller. Monocular pre-crash vehicle detection : Features and classifiers. *IEEE transactions on image processing*, 15(7) :2019–2034, 2006.

[Sem04] D. Semani. *Une méthode supervisée de sélection et de discrimination avec rejet. Application au projet Aquathèque.* PhD thesis, Université de La Rochelle, 2004.

[SGH04] Amnon Shashua, Yoram Gdalyahu, and Gaby Hayun. Pedestrian detection for driving assistance systems : Single-frame classification and system level performance. *IEEE Intelligent Vehicles Symposium*, pages 1–6, 2004.

[SGRB05] F. Suard, V. Guigue, A. Rakotomamonjy, and A. Benshrair. Pedestrian detection using stereo-vision and graph kernels. *IEEE Intelligent Vehicles Symposium*, pages 267–272, 2005.

[Sha78] G. Shafer. A mathematical theory of evidence. *Journal of the American Statistical Association*, 73(363) :677–678, 1978.

[SK94] P Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, 1994.

[SM97] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5) :530–534, 1997.

[Sme90] P. Smets. The combination of evidence in the transferable belief model. *IEEE transactions on pattern analysis and machine intelligence*, 12(5) :447–458, 1990.

[Sme93] P. Smets. Belief functions : The disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9(1) :1–35, 1993.

[SRBB06] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. *IEEE Intelligent Vehicles Symposium*, pages 206–212, 2006.

[Sua06] F. Suard. *Méthodes à noyaux pour la détection de piétons.* PhD thesis, l'Institut National des Sciences Appliquées de Rouen, 2006.

[SZ03] J. Sivic and A. Zisserman. Video google : A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[TBM+06] G. Toulminet, M. Bertozzi, S. Mousset, A. Bensrhair, and A. Broggi. Vehicle detection by means of stereo vision-based obstacles features extraction and monocular pattern analysis. *IEEE Transactions on Image Processing*, 15(8) :2364–2375, 2006.

[TLF08] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[TPM06] 0. Tuzel, F. Porikli, and P. Meer. Region covariance : A fast descriptor for detection and classification. *European Conference on Computer Vision*, pages 589–600, 2006.

[TPM08] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 30(10) :1713–1727, 2008.

[TS98] C. Tzomakas and W.V. Seelen. Vehicle detection in traffic scenes using shadows. Technical report, Institut für Neuroinformatik, Ruhr-Universität Bochum, 1998.

[Vap95] V.N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., 1995.

[VJ02] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2) :137–154, 2002.

[VJS03] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Conference on Computer Vision*, pages 734–741, 2003.

[WCG03] C. Wallraven, B. Caputo, and A.B.A. Graf. Recognition with local features : the kernel recipe. In *International Conference on Computer Vision*, pages 257–264, 2003.

[WN05] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. *IEEE International Conference on Computer Vision*, pages 90–97, 2005.

[WN06] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2006.

[WN07] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75 :247–266, 2007.

[WWP00] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *European Conference on Computer Vision*, pages 18–32, 2000.

[XLF] F. Xu, X. Liu, and K. Fujimura. Pedestrian detection and tracking with night vision. *IEEE Transactions on Intelligent Transportation Systems*, 6(1) :63–71.

[Yag87] R. Yager. On the dempster-shafer framework and new combinaison rules. *Informations Sciences*, 41 :93–137, 1987.

[YJZ10] I Ye, J Jiao, and B. Zhang. Fast pedestrian detection with multi-scale orientation features and two-stage classifiers. *International Conference on Image Processing*, pages 881–884, 2010.

[ZT00] L. Zhao and C.E Thorpe. Stereo- and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3) :148–154, 2000.

Liste des publications

B. Besbes, A. Rogozan, A. Bensrhair, *Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images*, IEEE Intelligent Vehicles Symposium (IV '10), pages 156 161, San Diego, 21-24 Juin 2010.

B. Besbes, B. Labbe, A. Bensrhair, A. Rogozan, *SVM-based fast pedestrian recognition using a hierarchical codebook of local features*, IEEE International Workshop on machine learning for signal processing (MLSP'10), pages 226-231, Finland , 29-1 September-Octobre 2010.

B. Besbes, A. Apatean, A. Rogozan, A. Bensrhair, *Combining SURF-based Local and Global features for Road Obstacle Recognition in Far Infrared Images*, IEEE Conference on Intelligent Transportation Systems (ITSC'10), pages 1869-1874, Portugal, 19-22 September 2010.

B. Besbes, S. Ammar, Y. Kessentini, A. Rogozan, A. Bensrhair, *Evidential combination of SVM road obstacle classifiers in visible and far infrared images*, IEEE Intelligent Vehicles Symposium (IV'11), pages 1074-1079, Baden-Baden, 5-9 Juin 2011.

B. Besbes, A. Rogozan, A. Bensrhair, *Vocabulaire Visuel Hiérarchique pour la détection et le suivi de piétons en utilisant l'infrarouge lointain*, XIIIe Colloque GRETSI - Traitement du Signal et des Images (GRETSI'11), Bordeaux, 5-8 septembre 2011

Résumé :
Cette thèse s'inscrit dans le contexte de la vision embarquée pour la détection et la reconnaissance d'obstacles routiers, en vue d'application d'assistance à la conduite automobile.
À l'issue d'une étude bibliographique, nous avons constaté que la problématique de détection d'obstacles routiers, notamment des piétons, à l'aide d'une caméra embarquée, ne peut être résolue convenablement sans recourir aux techniques de reconnaissance de catégories d'objets dans les images. Ainsi, une étude complète du processus de la reconnaissance est réalisée, couvrant les techniques de représentation, d'apprentissage et de fusion d'informations. Les contributions de cette thèse se déclinent principalement autour de ces trois axes.
Notre première contribution concerne la conception d'un modèle d'apparence locale basé sur un ensemble de descripteurs locaux SURF (Speeded Up Robust Features) représentés dans un Vocabulaire Visuel Hiérarchique. Bien que ce modèle soit robuste aux larges variations d'apparences et de formes intra-classe, il nécessite d'être couplé à une technique de classification permettant de discriminer et de catégoriser précisément les objets routiers. Une deuxième contribution présentée dans la thèse porte sur la combinaison du Vocabulaire Visuel Hiérarchique avec un classifieur SVM (Support Vecteur Machine).
Notre troisième contribution concerne l'étude de l'apport d'un module de fusion multimodale permettant d'envisager la combinaison des images visibles et infrarouges. Cette étude met en évidence de façon expérimentale la complémentarité des caractéristiques locales et globales ainsi que la modalité visible et celle infrarouge. Pour réduire la complexité du système, une stratégie de classification à deux niveaux de décision a été proposée. Cette stratégie est basée sur la théorie des fonctions de croyances et permet d'accélérer grandement le temps de prise de décision.
Une dernière contribution est une synthèse des précédentes : nous mettons à profit les résultats d'expérimentations et nous intégrons les éléments développés dans un système de détection et de suivi de piétons en infrarouge-lointain. Ce système a été validé sur différentes bases d'images et séquences routières en milieu urbain.

Mots-clés : Vision embarquée, Détection et reconnaissance d'obstacles routiers, Représentation des images, Classification par SVM, Fusion de capteurs, Fonction de croyances, Détection de piétons en Infrarouge-lointain.

Abstract:
The aim of this thesis arises in the context of Embedded-vision system for road obstacles detection and recognition: application to driver assistance systems.
Following a literature review, we found that the problem of road obstacle detection, especially pedestrians, by using an on-board camera, cannot be adequately resolved without resorting to object recognition techniques. Thus, a preliminary study of the recognition process is presented, including the techniques of image representation, Classification and information fusion. The contributions of this thesis are organized around these three axes.
Our first contribution is the design of a local appearance model based on SURF (Speeded Up Robust Features) features and represented in a hierarchical Codebook. This model shows considerable robustness with respect to significant intra-class variation of object appearance and shape. However, the price for this robustness typically is that it tends to produce a significant number of false positives. This proves the need for integration of discriminative techniques in order to accurately categorize road objects. A second contribution presented in this thesis focuses on the combination of the Hierarchical Codebook with an SVM classifier.
Our third contribution concerns the study of the implementation of a multimodal fusion module that combines information from visible and infrared spectrum. This study highlights and verifies experimentally the complementarities between the proposed local and global features, on the one hand, and visible and infrared spectrum on the other hand. In order to reduce the complexity of the overall system, a two-level classification strategy is proposed. This strategy, based on belief functions, enables to speed up the classification process without compromising the recognition performance.
A final contribution provides a synthesis across the previous ones and involves the implementation of a fast pedestrian detection system using a far-infrared camera. This system was validated with different urban road scenes that are recorded from an onboard camera.

Keywords: Embedded vision, Road obstacle detection and recognition, Image representation, SVM classification, Sensor fusion, Belief functions, Pedestrian detection in far-infrared images.