UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

RECONNAISSANCE D'OBJETS EN MOUVEMENT DANS LA VIDÉO PAR DESCRIPTION GÉOMÉTRIQUE ET APPRENTISSAGE SUPERVISÉ

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

YAOVI AHADJITSE

NOVEMBRE 2013

Ce mémoire a été évalué par un jury composé des personnes suivantes :

Dr. Mohand Saïd Allili ..... Directeur de recherche
Dr. Nadia Baaziz ..... President du jury
Dr. Ana-Maria Cretu ..... Membre du jury

Remerciements

Je tiens à remercier en premier lieu le Dr Mohand Saïd Allili, pour avoir accepté de m'encadrer et de me guider tout au long de ce mémoire. Je le remercie également pour m'avoir accueilli dans son laboratoire LARIVA.

Je remercie tout particulièrement ma famille pour son support constant durant ces années d'études.

Je remercie également les Dres Nadia Baaziz et Ana-Maria Cretu pour avoir accepté d'évaluer ce travail en leur qualité de membre du jury.

Je remercie enfin mes camarades du laboratoire LARIVA avec lesquels j'ai eu des échanges enrichissants.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations, sigles et acronymes

BoF Bag-of-Features

FC Forme Contextuelle

HOG Histogram of Oriented Gradient

KPPV K Plus Proches Voisins

SURF Speeded Up Robust Feature

SUZAN Smallest Univalue Segment Assimilating Nucleus

SVH Système Visuel Humain

Résumé

La reconnaissance des objets en mouvement dans les vidéos est un problème important en vision artificielle et en traitement d'images. Cette tâche est très utile vue l'accroissement du nombre de vidéos générées par les médias numériques (ex., internet, la télévision, les vidéos personnelles, la surveillance vidéo). La reconnaissance automatique des objets en vidéos peut ainsi renforcer la sécurité, faciliter la gestion des vidéos ainsi que permettre de nouvelles applications en interaction personne/machine. Les méthodes existantes dans la littérature utilisent certaines approches par description géométrique et proposent des solutions pour accomplir cette tâche. Cependant, ces solutions demeurent loin de rivaliser la capacité du système de la vision humaine.

Dans le cadre de ce travail, nous proposons une nouvelle approche de reconnaissance des objets en mouvement dans les vidéos. Cette approche se base sur la création de descripteurs tenant compte de la géométrie globale et locale des objets et permettant de représenter de façon unique la forme des objets. Notre approche se base également sur l'utilisation de méthodes d'apprentissage statistique supervisé et d'ontologie de formes, permettant de reconnaître automatiquement les catégories d'objets ainsi que leurs poses. Des expérimentations sur plusieurs types d'objets ont permis de valider la performance de notre approche.

Abstract

The recognition of moving objects in video is an important area of research in computer vision and in image processing. This task is very useful to manage the increasing number of videos generated by digital media (eg., Internet , television, home video , video surveillance). The automatic recognition of objects in videos can enhance safety, facilitate the video management and lead to new applications based on human/computer interaction.

The existing methods in the literature tend to identify objects by creating descriptors based on their geometrical properties. However, these solutions are still far from rivaling the excellent capabilities of the human vision system.

In this work, we propose a new approach for the recognition of moving objects in videos. This approach is based on creating descriptors that aim to accurately represent the local and global geometrical properties of objects and therefore, allowing a unique description of objects' shapes. Our approach uses supervised machine learning methods in order to automatically recognize moving objects. We've also built an ontology which represents different types of objects and their poses. Experiments with these objects have validated the performance of our approach.

Chapitre 1

Introduction

1.1 Généralités

Les images et les vidéos occupent une place de plus en plus importante dans la société contemporaine. Les médias numériques (par ex., la télévision, l'internet, les studios de production cinématographique, la surveillance vidéo, les vidéos personnelles) génèrent une quantité très importante d'images numériques et de vidéos. La baisse des coûts de matériels d'acquisition, de stockage et les nouvelles techniques de compressions ont favorisé la création à grande échelle de ces derniers. Par ailleurs, les images numériques et la vidéo sont devenues indispensables pour divers domaines d'application, tels que la détection d'intrusions pour la sécurité, la surveillance du trafic routier, la médecine pour l'imagerie médicale, ou encore lors des événements sportifs (ex., renforcement de l'arbitrage, création automatique de résumés).

Des contraintes d'exploitation découlent des observations citées ci-dessus, parmi lesquelles nous citerons celles qui sont liées à la reconnaissance des objets en mouvement dans les vidéos. Par exemple, de nos jours, un très grand nombre de caméras est déployé exclusivement pour la surveillance vidéo. Souvent, le contenu de ces vidéos est interprété par des opérateurs humains qui engendrent des coûts exorbitants pour le suivi et l'analyse du contenu, sans mentionner les erreurs qui peuvent être induites par la fatigue et

l'inattention humaine. Un des problèmes importants abordés dans la surveillance vidéo est la reconnaissance des types d'objets en mouvement et leurs actions, afin de détecter, par exemple, des menaces potentielles (ex., vols, attentats, accidents), ou tout simplement pour des fins de statistiques (ex., compter le nombre d'individus, de voitures dans une entrée de parc)

La reconnaissance des objets peut être aussi très utile pour construire des bases de données indexées en termes de contenu vidéos [6]. Il serait, par exemple, utile d'envoyer des requêtes spécifiques à ces dernières pour récupérer des scènes contenant certains types d'objets [47]. Des techniques d'indexation connues comme l'utilisation de systèmes de *cluster* [59] aident actuellement dans le parcours rapide des scènes dans les vidéos.

On peut citer aussi la création de sommaires de vidéos qui peuvent permettre aux utilisateurs de retrouver des scènes spécifiques afin d'éviter de parcourir la vidéo en entier. L'une des techniques les plus utilisées pour la création de résumé de scènes utilise les *trames clés* de la vidéo [6, 60]. Il s'agit de créer des sommaires de plusieurs trames ayant un contenu visuel semblable en calculant une trame moyenne, médiane, etc. Cependant, ces trames, ayant été calculées à partir d'information visuelle globale, elles ne contiennent pas d'information précise sur les types d'objets en mouvement présents dans les scènes. Dans le cas de la surveillance vidéo où les vidéos acquises peuvent être très longues à visionner, il serait souhaitable de pouvoir localiser les segments ou les scènes contenant des objets spécifiques dans des intervalles de temps limités. Récemment, plusieurs progrès de recherche ont été réalisés pour la reconnaissance automatique des objets (ex., faces, humains [14, 55]) dans les images et les vidéos. Cependant, la plupart de ces méthodes ne sont pas facilement extensibles pour différents types d'objets et, souvent, n'atteignent pas la précision souhaitée qui serait équivalente à celle de l'humain.

Le système visuel de l'humain (SVH) est capable de reconnaître facilement les objets dans les images et les vidéos, même lorsque les conditions de vue sont mauvaises (ex., changement de points de vue, de taille d'objets). Cette tâche s'effectue instantanément et sans effort perceptible du SVH. Ce processus de reconnaissance fait intervenir différentes parties du cerveau humain qui interprètent les signaux captés au niveau de la rétine des yeux. Les études de *Selfridge* [45, 51] sur le *pandemonium* ont montré que le SVH se sert du concept de descripteurs pour la reconnaissance des objets. Ces descripteurs sont utilisés dans le processus de bas niveau pour la reconnaissance automatique des objets. Dans ses travaux, *David Marr* [42] a montré aussi comment le SVH représente de façon progressive les objets à partir de leurs descripteurs. De plus, la vision humaine s'accompagne de fonctionnalités sémantiques qui aident dans la reconnaissance des objets dans leurs contextes [22]. Récemment, des systèmes intelligents d'analyse de contenu de vidéos ont été créés, tels que celui de *Google* [57]. Ce dernier utilise des descripteurs pour isoler des objets statiques ou en mouvement dans des vidéos. Le *Smart Surveillance System (SS3)* de *IBM* [54] est un autre type de système intelligent qui peut accepter des alertes d'événements à travers des interfaces, et qui peut fournir certaines statistiques sur le contenu de la vidéo. Malgré les progrès importants réalisés dans le domaine de la vision artificielle, beaucoup de travail reste encore à accomplir pour une reconnaissance efficace des objets qui pourrait rivaliser celle du système visuel humain.

Il est enfin utile de mentionner que la reconnaissance des objets par les humains est facilitée par la performance du fonctionnement du cortex visuel au niveau du cerveau [49]. Ce mécanisme de fonctionnement permet aux humains de disposer d'une base de connaissances leur permettant d'analyser les objets et de les identifier. Malheureusement, les ordinateurs de nos jours qui sont utilisés dans la vision artificielle ne disposent pas d'une

base de connaissances équivalente à celle du SVH, qui pourrait caractériser chaque type d'objet et permettre sa reconnaissance de manière automatique. Il y va, donc, de soi que tout algorithme qui peut être développé pour la reconnaissance automatique des objets doit tenir compte de cette propriété, en construisant une base d'apprentissage facilitant la reconnaissance des objets. En d'autres termes, il est nécessaire de procéder d'abord par une phase d'apprentissage qui doit extraire l'information pertinente pour caractériser chaque type d'objet pour permettre d'identifier ultérieurement ses instances dans de nouvelles images et vidéos.

1.2 Contexte de notre recherche

La reconnaissance des objets par la vision artificielle est un domaine de recherche qui s'étend actuellement à diverses branches de l'industrie. On peut citer par exemple l'implémentation des systèmes de vision artificielle dans les nouvelles générations de voitures et également dans les *smartphones*. D'un côté, nous avons les applications traditionnelles pour ce type de système dans l'objectif d'effectuer des recherches d'objets dans les bases de données d'images comme les applications *CBIR (Content Based Image Retrieval)* [6]. De l'autre, il faudra remarquer le besoin grandissant d'implémenter ces systèmes de reconnaissance des objets pour la grande quantité des vidéos qui sont présentement générées. Ces systèmes profiteront également des évolutions technologiques dans les domaines de la compression et du codage des données. L'efficacité de la reconnaissance des objets est étroitement liée à celle des descripteurs qui sont utilisés pour les représenter dans les systèmes de traitement. Les recherches qui sont actuellement faites dans le domaine de la vision artificielle ont entre autres pour objectif de trouver le

descripteur idéal pour la plupart des objets connus ; en essayant de répondre aux questions suivantes : quel est le meilleur descripteur capable d'identifier la plupart des objets existants ? Quelle est la quantité d'informations nécessaires pour décrire efficacement ces objets ?

Le travail de ce mémoire se situe dans un contexte d'évolutions technologiques liées aux domaines de la vision artificielle et du traitement de la vidéo avec un besoin de trouver des méthodes permettant un parcours rapide et une reconnaissance des objets en mouvement présents dans les vidéos. Les applications potentielles de ce travail sont multiples, telles que la reconnaissance d'objets dans les vidéos de surveillance, la création de nouveaux types de sommaires de vidéos à partir des objets en mouvement dans le but de faciliter un parcours rapide des vidéos.

1.3 Problématique de notre recherche

La reconnaissance des objets est une tâche simple et triviale pour les humains. Le SVH est capable de faire la distinction, d'une part, entre des objets et l'arrière-plan d'une image et d'autre part, entre plusieurs objets présents dans une scène de vidéo. Dans le cas de la vision par ordinateur, le processus est beaucoup plus difficile car on ne dispose pas des mêmes mécanismes de mémoire, d'apprentissage et de traitement qui caractérisent le SVH. Cependant, les ordinateurs de nos jours possèdent une grande capacité de stockage et de traitement qui peuvent être exploitées pour construire des algorithmes puissants pour la détection et la reconnaissance automatiques des objets.

Récemment, plusieurs approches ont été proposées pour la reconnaissance des objets. La plupart de ces dernières sont basées sur le développement de descripteurs permet-

tant de représenter les objets et sur l'utilisation d'algorithmes d'apprentissage pour les identifier. Pour une reconnaissance efficace des objets, un descripteur doit pouvoir les représenter de façon unique et fiable. Par ailleurs, les objets peuvent être représentés et identifiés selon leurs caractéristiques globales (ex., contours, silhouettes) ou locales (ex., points d'intérêts) [34]. Chacune de ces caractéristiques possède des avantages et des inconvénients. Les caractéristiques globales permettent de décrire la forme globale des objets (la forme de leur silhouette), mais elles sont sensibles aux déformations de ces derniers. Dans le cas des silhouettes humaines, par exemple, une personne peut s'incliner ou se déplacer avec différentes allures, cela entraîne une très grande variabilité dans la forme des silhouettes. Les caractéristiques locales permettent d'identifier les objets sans extraire leur contour [24].

Dans la vidéo, la détection et la reconnaissance des objets peuvent être facilitées par la présence du mouvement. En effet, l'analyse des vidéos peut se focaliser uniquement sur celles de régions présentant du mouvement, et ainsi réduire considérablement le temps de traitement. Généralement, des techniques de soustraction de fond sont utilisées pour isoler les objets en mouvement dans la vidéo [11, 33, 50]. Cependant ces techniques peuvent produire des silhouettes dépourvues de sémantique qui doivent être raffinées pour être correctement identifiées. Par exemple, un groupe de personnes en mouvement peut être détecté comme une seule silhouette et identifiée comme une seule personne. Une solution possible serait alors de disposer d'un ensemble d'images assez variées permettant de prendre en compte la majorité des cas qui peuvent se présenter. Cependant, cette solution engendre un autre problème qui est lié au temps de traitement des algorithmes de reconnaissance d'objets. Il s'agit d'ailleurs d'une contrainte essentielle surtout pour les systèmes en temps réel, tels que la vidéo surveillance. Contrairement aux humains,

les descripteurs d'objets sont représentés dans la machine sous forme de vecteurs. Ces vecteurs constituent la base de connaissances disponible pour identifier les objets. Les algorithmes de reconnaissance d'objets ont donc pour principale tâche de construire des descripteurs assez fiables pour représenter les objets et les reconnaître de manière efficace et automatique.

Partant de ces remarques, un des problèmes importants en vision artificielle demeure la construction de bases de connaissances adéquates offrant une représentation sémantique pouvant faciliter la reconnaissance automatique des objets. Dans les vidéos par exemple, le problème est notamment lié à l'extraction des objets en mouvement dans les scènes, et aussi à leur reconnaissance par rapport à différentes classes d'objets.

1.4 Objectifs de notre travail

Nous nous proposons dans ce mémoire d'aborder la reconnaissance des objets en mouvement dans les vidéos en utilisant une approche basée sur l'analyse des silhouettes et l'utilisation de méthodes d'apprentissage supervisé. Les objectifs de notre travail regroupent les points suivants :

1. Développement des descripteurs adaptés pour l'analyse des silhouettes d'objets afin de faciliter leur reconnaissance. Ces descripteurs doivent répondre aux contraintes d'invariance telles que l'invariance à l'échelle ou à la translation d'objets. Dans une première phase, nous utiliserons des descripteurs basés sur les contours d'objets. Dans une seconde phase, nous développerons des descripteurs basés sur les points d'intérêt.

2. L'utilisation de différentes méthodes de classification pour la reconnaissance des objets. Pour les descripteurs basés sur les contours d'objets, nous utiliserons la méthode des *K Plus Proches Voisins* (KPPV) qui est une méthode de classification non-paramétrique, et le classificateur naïf de Bayes qui tient compte des distributions des classes d'objets. Pour les descripteurs basés sur les points d'intérêt, nous utiliserons les méthodes de mise en correspondance.

3. Nous construirons une base de données annotée pour l'apprentissage et pour les tests de reconnaissance d'objets. Les résultats concluants de ces tests pourront conduire à un déploiement de notre système dans les domaines comme celui de la surveillance de vidéos. Des métriques connues pour mesurer les erreurs de classification seront utilisées pour évaluer la qualité de la reconnaissance des objets par notre approche.

Chapitre 2

État des connaissances

2.1 Introduction

Le processus de reconnaissance des objets dans le domaine de la vision artificielle se déroule en général selon les étapes suivantes :

- la représentation des objets selon les caractéristiques de bas niveau de l'image (ex., couleur, texture, gradient d'intensité)
- la classification des caractéristiques d'objets inconnus par rapport à des caractéristiques d'objets existants [35, 61, 63].

Il s'agit de l'une des tâches les plus complexes que les machines sont amenées à résoudre dans la vision artificielle. Il va de soi que les objets dans les images sont des entités statiques, contrairement à la vidéo où l'information du mouvement peut être utilisée pour extraire les objets en mouvement. La reconnaissance des objets peut ainsi être vue selon deux axes. Le premier axe est relatif à la reconnaissance des objets sur les images et le second à la reconnaissance dans les vidéos. Différentes méthodes ont été développées pour la reconnaissance des objets dans les deux axes d'analyse, bien que certaines méthodes qui sont développées pour les images sont aussi valables pour les vidéos.

2.2 Reconnaissance d'objets dans les images

Nous désignerons dans l'ensemble de ce document, une image par le terme $I(X)$. Dans le cas du traitement de la vidéo, le terme $I_t(X)$ sera utilisé pour désigner une image à un instant t dans le domaine temporel, avec $t \in 1, 2, 3, \ldots$. Le terme $X = (x, y)$ désignera les coordonnées des pixels dans le domaine spatial. La reconnaissance des objets dans les images se déroule à travers un processus dont une première étape est la segmentation.

2.2.1 La segmentation d'images

Elle consiste à partitionner une image en régions distinctes ou objets [21, 62]. La taille des différentes régions varie en fonction des exigences, du niveau de finesse requis et également du contexte d'utilisation ou de l'application sous-jacente [53]. Si nous supposons que R définit la région entière d'une image, la segmentation se définira formellement comme la subdivision de R en sous-régions $R_1, R_2, \ldots, R_n$, respectant les conditions suivantes :

1. $R = \bigcup_{i=1}^{n} R_i$.
2. R_i forme un ensemble de pixels connectés avec $i = 1, 2, \ldots, n$.
3. $R_i \cap R_j = \emptyset \ \forall i, j$ telle que $, i \neq j$.
4. $Q(R_i) = VRAI,$ pour $i = 1, 2, \ldots, n$.
5. $Q(R_i \cup R_j) = FAUX$ pour deux régions adjacentes R_i et R_j.

La première condition indique que tout pixel doit nécessairement appartenir à une des régions. La seconde condition impose que les pixels d'une région soient connectés selon un mode défini (4-connexité ou 8-connexité) [21]. La troisième condition indique

que les régions doivent être disjointes. La quatrième condition stipule le prédicat que doivent satisfaire les pixels appartenant à une région. Le prédicat $Q(R_i) = VRAI$ peut indiquer que tous les pixels ont la même intensité. Enfin, la dernière condition exprime le fait que deux régions adjacentes, unies, ne doivent pas satisfaire les prédicats qui sont définis par la quatrième condition.

Dans les images, les objets sont statiques et apparaissent sur un arrière-plan qui peut être uniforme ou encombré. Certaines images contiennent plusieurs objets dont les contours ne sont pas tout le temps bien définis. La première étape dans le processus de reconnaissance des objets dans les images consiste à rechercher les contours ou régions qui les composent puis à les segmenter [22, 30, 32]. Les contours peuvent être extraits par des algorithmes de détection de contours (ex. *canny edge detector*) [21]. Les algorithmes de segmentation se basent généralement sur les propriétés qui sont relatives, soit à la *discontinuité*, soit à la *similarité* des valeurs des pixels de l'image [21]. Concernant la discontinuité, l'image est partitionnée en fonction des changements brusques des valeurs des pixels ; pour la similarité, l'image est partitionnée en fonction d'un ensemble de critères prédéfinis permettant de regrouper les pixels en régions homogènes (ex., seuillage, variance d'intensité) [21].

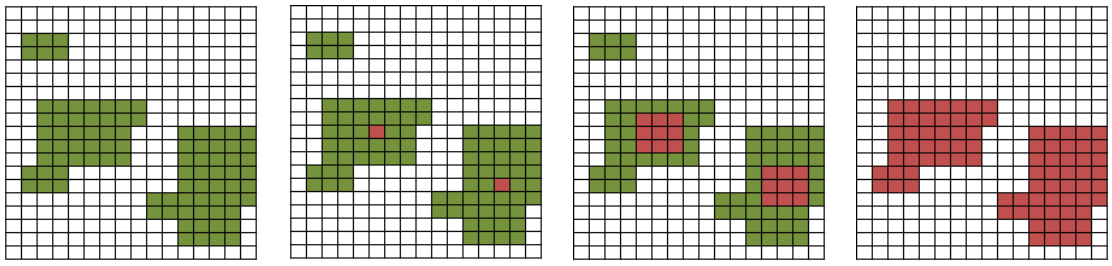
La segmentation d'images peut être effectuée par différents algorithmes [53, 62]. Nous présentons dans la section suivante quelques techniques de segmentation utilisées dans les images.

2.2.1.1 Méthodes par croissance de régions

La segmentation d'images par croissance de régions consiste à regrouper des pixels d'une image en régions, sur la base d'un critère d'appartenance ou prédicat d'homogé-

néité [21]. Le critère d'appartenance peut être, par exemple, la moyenne ou la variance des intensités des pixels.

La croissance de régions débute par un ou plusieurs pixels à partir desquels les régions vont croître progressivement. Le choix de pixels de départ, encore connus sous l'appellation *germes* ou *semences*, peut s'effectuer de façon automatique ou de façon manuelle par un humain. Les pixels qui se trouvent dans les voisinages des *germes* et qui respectent le prédicat d'homogénéité permettent d'élargir ces *germes* pour constituer les différentes régions de l'image. Dans l'absence d'un critère prédéfini, la procédure consiste à calculer pour les pixels de l'image un ensemble de propriétés qui permettront de les regrouper par similarité. La figure 2.1 montre une illustration des étapes de la segmentation par croissance de régions. La figure 2.1(a) montre l'image originale à segmenter avec trois régions différentes. Sur la figure 2.1(b), on remarque les deux germes initiaux (en rouge) qui sont les points de départ du processus de croissance des régions. La figure 2.1(c) donne une illustration de la croissance des germes et la figure 2.1(d) montre l'image segmentée.



(a) Image à segmenter (b) Germes initiaux en rouge (c) Croissance des germes (d) Image segmentée

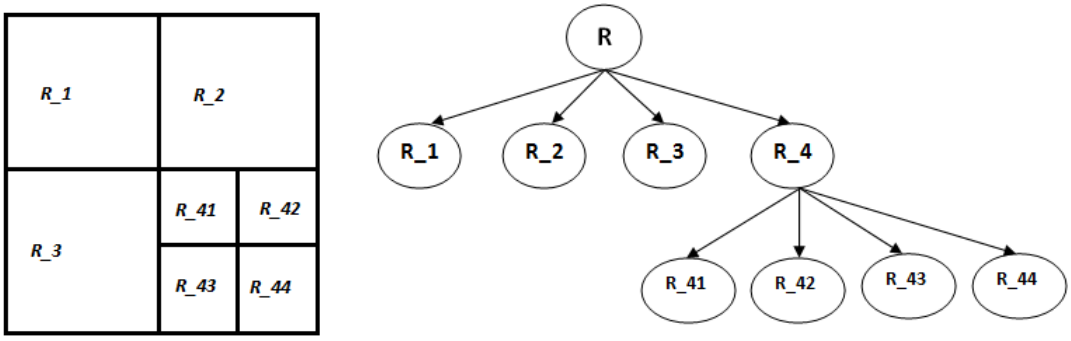
FIGURE 2.1 – Exemple de segmentation par croissance de régions.

D'autres critères additionnels sont définis pour l'arrêt du processus de croissance des régions. Il peut s'agir par exemple de la taille des régions, de la forme des régions, ou encore de la comparaison entre l'intensité d'un pixel avec la moyenne des pixels de la région.

2.2.1.2 Méthodes par division-fusion de régions

Cette méthode de segmentation se déroule en deux étapes : dans la première étape l'image est divisée en un certain nombre arbitraire de régions disjointes, et dans la seconde étape les différentes régions sont regroupées selon un prédicat. Soient R la région entière de l'image, Q le prédicat d'homogénéité devant être satisfait par chaque région. Ce prédicat peut également être relatif à la moyenne ou à la variance des intensités des pixels. L'opération de la première étape divise l'image de façon récursive en quadrants ; chaque quadrant qui satisfait les conditions du prédicat Q est considéré comme région homogène. Chaque fois que le prédicat n'est pas vérifié pour un quadrant, ce dernier est subdivisé de nouveau en d'autres quadrants. À la fin de ce processus, une structure arborescente des quadrants (*quadtrees*) est construite. Chaque noeud racine de la structure arborescente correspond à une subdivision en quatre sous-régions. La seconde étape du processus effectue l'opération dans le sens contraire et fusionne des régions adjacentes en d'autres régions de plus grande taille. La fusion des régions s'effectue également en fonction du même prédicat Q. Ainsi deux régions adjacentes R_j et R_k peuvent être fusionnées si et seulement si elles satisfont la condition $Q(R_j \cup R_k) = VRAI$.

Les figures 2.2(a) et 2.2(b) montrent respectivement la division en quadrants avec les différentes régions, et la structure arborescente résultante.



(a) Division en quadrants (b) Structure arborescente de la figure 2.2(a)

FIGURE 2.2 – Exemple de segmentation par division-fusion [21].

Des travaux antérieurs ont utilisé des méthodes de segmentation pour la détection et la reconnaissance des objets dans les images. Oren *et al.* [38], par exemple, ont utilisé les propriétés des ondelettes pour la segmentation d'images en régions en vue de la reconnaissance des objets. Les régions correspondent à des modèles de tailles fixes de (ex. 16×16 ou 32×32 pixels) ; un objet pouvant être composé de plusieurs régions. Les différentes régions homogènes sont comparées avec une base de données d'exemples dans le but d'identifier des silhouettes humaines. Gu *et al.* [22] ont utilisé la segmentation par régions pour la reconnaissance des objets. Pour chaque objet, un modèle est construit en le décomposant en plusieurs parties représentées dans une structure arborescente.

2.2.1.3 Méthodes par agrégation

D'autres méthodes existent pour la segmentation en se basant sur l'agrégation des pixels par des techniques de groupement. Trois méthodes peuvent être invoquées dans cette catégorie [29] :

– **Méthode du mean-shift [12]**

Cette méthode de segmentation permet de créer des agrégats des pixels d'une image. Ces agrégats sont constitués en fonction des différentes caractéristiques des pixels de l'image. La formation de ces agrégats ne suppose pas de connaissance a priori sur la distribution des pixels dans l'image. Les agrégats indiquent la présence de pixels ayant les mêmes caractéristiques et qui constitueront les régions après segmentation. Différentes méthodes sont utilisées pour la création de l'agrégat dont nous citerons celle du noyau pour le calcul de la densité de probabilité. Un pixel de l'image est affecté à un agrégat en fonction de son voisinage. Il sera ainsi affecté à l'agrégat ayant la plus grande densité dans son voisinage et qui indique la direction de l'agrégat. Cette direction dépend de la première dérivée d'un noyau ou du gradient de la distribution de ses pixels. Ainsi, les distributions de pixels ayant une certaine uniformité indiquent des régions à faible gradient. Les pixels qui sont situés à la frontière des régions sont ceux dont le gradient est élevé. La formation d'un agrégat est un processus itératif. Ce processus débute par un point de départ X_d quelconque dans l'image et par des calculs successifs de vecteurs *mean-shifts* à partir des autres points X_v dans son voisinage. Ces calculs s'effectuent par la formule suivante :

$$m_h(X_v) = \frac{\sum_{i=1}^{N} X_d g(|\frac{(X_v - X_d)}{N}|^2)}{\sum_{i=1}^{N} g(|\frac{(X_v - X_d)}{N}|^2)} - X_d, \tag{2.1}$$

dans laquelle g représente le gradient d'un point et $m_h(X_v)$ représente le vecteur *mean-shift* calculé entre les points X_d et X_v.

– **Méthode des K-moyennes**

Il s'agit d'une méthode de segmentation dont le principe repose sur la formation d'agrégats des pixels de l'image [5]. Cette méthode permet de fixer à l'avance le nombre d'agrégats à constituer. Elle considère des critères d'homogénéité des pixels de l'image. Ces critères peuvent être par exemple la texture ou le niveau de gris des pixels de l'image. Cette méthode de segmentation calcule les moyennes des différents agrégats qui sont constitués. Ces moyennes représentent les centres des agrégats et leur calcul s'effectue de façon itérative. À chaque itération, chaque agrégat est enrichi avec les pixels qui sont les plus proches et la moyenne de l'agrégat est mise à jour. Le processus de segmentation s'arrête en fonction d'un critère qui peut être relatif à la stabilité des centres des agrégats qui sont constitués.

– **Méthode par coupure de graphes**

Dans cette méthode, l'image est représentée par un graphe dont les noeuds constituent les pixels de l'image et les arêtes indiquent les liens entre un pixel et ses voisins [7]. La méthode par coupure de graphe associe des poids aux arêtes en fonction du niveau de gris des pixels qui sont liés. Le poids associé à une arête est faible lorsqu'il existe une grande différence entre les niveaux de gris des pixels liés. Ce poids est en revanche élevé dans le cas contraire. Cette méthode essaie de réaliser une *coupure* dans le graphe représentant l'image en fonction du *coût* des arêtes (poids associés) et d'un critère relatif au contour des régions à segmenter. L'objectif final est d'optimiser le coût total engendré. Pour effectuer la segmentation, l'utilisateur choisit au départ deux *germes* (pixels) dont l'un est considéré comme pixel du fond (*sink*) et l'autre comme appartenant à un objet (*source*).

La coupure dans le graphe s'effectue en cherchant le flot maximum entre les deux germes (*source* et *sink*). La figure 2.3 montre une illustration du processus.

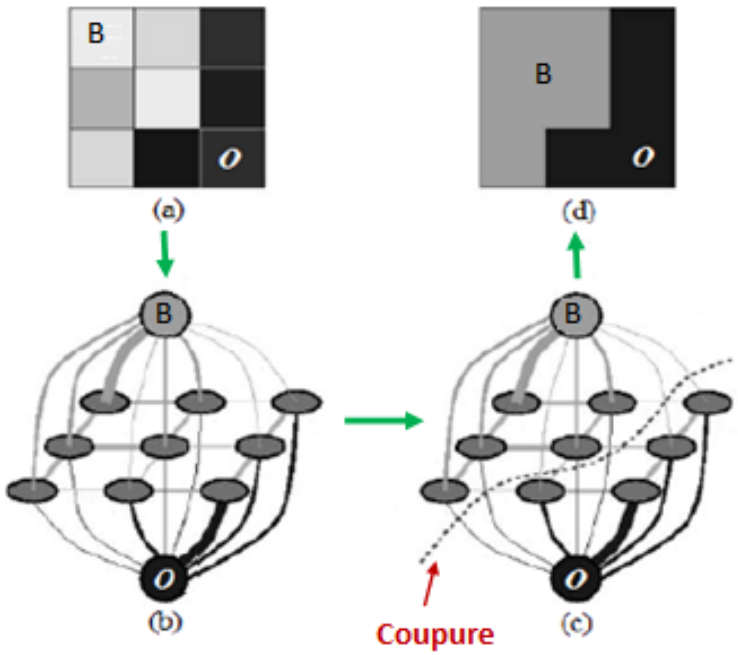
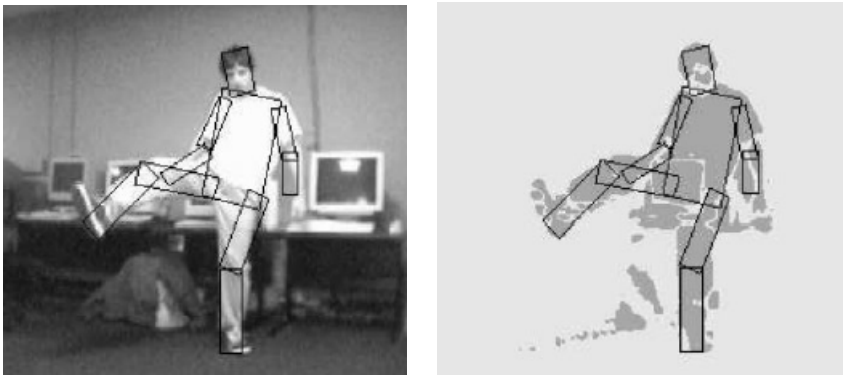


FIGURE 2.3 – Segmentation par coupure de graphe [7].
(a) Image avec deux germes initiaux : O=source et B=fond, (b) graphe, (c) coupure, (d) image segmentée.

2.2.1.4 Reconnaissance par la segmentation

La reconnaissance des objets dans les images peut également se faire par une modélisation des différentes parties de l'objet et par l'utilisation d'algorithmes appropriés permettant de les recomposer. Dans les travaux de Felzenszwalb *et al.* [20] un modèle graphique sous la forme d'une structure arborescente a été créé représentant une silhouette humaine. La structure arborescente modélise par un ensemble de rectangles connectés,

différentes parties du corps (torse, la tête et les membres). La topologie de cette structure arborescente montre les positions et les orientations relatives des rectangles. Il faut noter, cependant, que les difficultés de cette approche concernent la spécification des relations géométriques entre les parties de l'objet, la création des descripteurs et l'algorithme de reconnaissance approprié. Sur la figure 2.4(a), le modèle graphique a été superposé sur une image binaire avec arrière-plan. Ce même modèle a été superposé sur l'objet de la figure 2.4(b) après l'élimination de l'arrière-plan.



(a) Modèle graphique sur un objet avec arrière-plan

(b) Modèle graphique sur un objet sans arrière-plan

FIGURE 2.4 – Reconnaissance d'objets par utilisation de modèle graphique [20].

Une autre manière de modéliser les images ou les objets afin de permettre leur reconnaissance consiste à utiliser une technique analogue à celle des *Bag-of-Word* ou *sacs de mots*. Dans le domaine de la classification des textes, cette technique consiste à identifier chaque document en fonction des mots qui lui sont caractéristiques. Chaque document est représenté par un histogramme de fréquence d'apparition des mots qu'il contient. Ce concept a été proposé dans [27] pour la classification de documents textuels. Dans le domaine de la vision par ordinateur, ce même concept a été utilisé pour la reconnaissance

d'objets. Cependant le terme *Bag-of-Word* plus approprié aux documents est dans ce cas remplacé par le terme *Bag-of-Visual-Words* ou *sacs de mots visuels* ou encore *Bag of Features (BoF)*. Les *mots visuels* sont constitués de descripteurs extraits dans différentes parties de l'image, à partir desquels un dictionnaire est construit par utilisation d'une méthode de partitionnement. Ce dictionnaire de mots visuels permet de générer des histogrammes de fréquences pour les descripteurs, qui forment les modèles de représentation de chaque objet à identifier. Les *sacs de mots visuels* ont été utilisés par Yang *et al.* [58] dans leurs travaux de classification de scènes de vidéos. La même technique a été utilisée par Wang *et al.* [56] pour la reconnaissance et la localisation d'objets dans les images.

Finalement, il est important de mentionner que l'arrière-plan des images peut contenir des informations utiles et discriminantes pouvant aider dans le processus de reconnaissance d'objets. Ces informations sont parfois qualifiées *d'informations contextuelles*. Le travail de Divvala *et al.* [15], par exemple, a mis l'accent sur l'intégration de l'information contextuelle dans le processus de reconnaissance d'objets. Cette méthode a permis d'améliorer les résultats de la reconnaissance des objets. Allili *et al.* [1] ont utilisé l'information contextuelle pour construire un indice de pertinence pour les caractéristiques des objets utilisés durant la segmentation.

2.2.2 La détection de contours d'objets

La détection de contours permet de mettre en évidence les discontinuités d'intensité ou de couleur dans les images. Ces discontinuités peuvent provenir d'un vrai contour d'objet ou d'un bruit [21]. Par conséquent, la plupart des algorithmes existants pour la détection de contours produisent des contours discontinus d'objets. Dans le passé,

plusieurs approches ont proposé de relier les contours détectés pour reconstruire les frontières d'objets connus [10]. D'autres méthodes ont utilisé les modèles déformables pour suivre les frontières d'un objet [39, 13]. Paragios *et al.* [39, 44] ont proposé des approches de segmentation basées sur les contours actifs pour les objets dont la forme est déjà connue. Dans cette approche l'apprentissage de la forme est en général fait à partir d'échantillons d'objets de même forme que l'objet à segmenter. Des contraintes sont définies sur l'évolution d'une fonction *level set* dans le but de s'aligner avec la frontière de l'objet sur une nouvelle image. Récemment, la méthode basée sur l'*histogramme des gradients orientés* (HoG) [14] a été proposée pour la détection d'objets [25]. Cette méthode calcule localement sur des *patchs* des descripteurs basés sur l'orientation des contours. En utilisant une méthode d'apprentissage supervisé SVM [53], le HoG a été appliqué avec succès pour la détection d'humains [14].

2.3 Reconnaissance d'objets dans la vidéo

Deux approches principales ont été utilisées pour reconnaître les objets dans la vidéo. La première se base sur l'extraction d'objets en mouvement suivie par la reconnaissance de leur classe. La deuxième se base sur la reconnaissance directe des objets en utilisant le niveau de gris (c.-à-d. sans détecter au préalable le mouvement).

2.3.1 Reconnaissance avec détection d'objets en mouvement

Dans cette section, nous utiliserons la terminologie suivante :

Segmentation du mouvement : elle consiste à marquer les pixels qui sont associés à des régions en mouvement dans les scènes de la vidéo. Dans certaines situations,

ces régions peuvent ne pas correspondre à des objets identifiables et seront donc dépourvues d'une sémantique bien définie.

Détection du changement : Elle fait référence aux méthodes qui sont utilisées pour détecter les pixels d'images qui sont caractéristiques de la présence de mouvement. Il s'agit d'un cas particulier de segmentation du mouvement qui prend en considération deux types de régions ; dans le cas des cameras statiques on parle de régions avec présence ou non du mouvement, dans le cas des cameras mobiles on parle de régions avec mouvement local ou global.

Modélisation de l'arrière-plan : Par cette technique, des méthodes permettent de créer un modèle d'arrière-plan qui sera utilisé pour détecter les objets en mouvement. Les modèles d'arrière-plan peuvent être créés par application de méthodes de différenciation ou de méthodes statistiques sur l'information contenue dans les pixels à différents instants d'une séquence vidéo.

Soustraction de fond : Cette opération permet d'isoler les pixels faisant partie d'un objet en mouvement de ceux appartenant au modèle d'arrière-plan.

La détection des objets en mouvement s'effectue généralement par l'utilisation des techniques relatives à la détection de changements dans les scènes de vidéos et aussi par les méthodes d'estimation du mouvement [53]. Contrairement à l'image, dans laquelle les objets sont statiques, la vidéo peut contenir des objets statiques ou en mouvement. Lorsqu'il y a occurrence de mouvement, la reconnaissance des objets peut tirer profit de cette information. Ainsi, il va de soi que l'efficacité de la reconnaissance des objets dépendra intimement de la qualité de la segmentation du mouvement qui sera opérée sur la vidéo. La segmentation permet de partitionner la vidéo en exploitant l'informa-

tion spatiale (ex. couleur, texture, forme, etc.), l'information temporelle (ex. flux de mouvement, détection de changement) ou la combinaison des deux [2, 6]. La complexité des scènes de la vidéo en termes de mouvement, de couleur, de texture ou de contraste détermine également le choix des techniques utilisées pour la détection des objets en mouvement [6].

Les méthodes de détection d'objets en mouvement peuvent être regroupées en deux parties : les méthodes de détection sans modélisation de l'arrière-plan et celles avec modélisation de l'arrière-plan.

2.3.1.1 Détection sans modélisation de l'arrière-plan

Les techniques de détection sans modélisation de l'arrière-plan consistent en général à effectuer une opération de *différence spatio-temporelle* des valeurs d'intensité des pixels qui constituent les trames de la vidéo [6]. Dans sa forme la plus simple elle se limite à utiliser les pixels de deux trames consécutives de la vidéo : une première trame considérée comme modèle d'arrière-plan et une seconde trame contenant des pixels pouvant refléter la présence de mouvement. Cette *différence spatio-temporelle* se définit formellement par l'équation 2.2.

$$I_{t,k} = I_t(X) - I_k(X), \tag{2.2}$$

dans laquelle $I_t(X)$ et $I_k(X)$ représentent respectivement les intensités des pixels des trames de la vidéo aux temps t et k. En général, k est la trame au temps $t = 0$ qui représente la scène dépourvue d'objets en mouvement.

Une technique de seuillage est utilisée afin d'éliminer les bruits ou les pixels qui ne reflètent pas la présence de mouvement. Elle peut être définie selon l'équation 2.3 dans laquelle $z_{t,k}$ représente le label des pixels et δ représente le seuil.

$$z_{t,k} = \begin{cases} 1 & \text{si } |I_{t,k}| > \delta \\ 0 & \text{sinon} \end{cases} \tag{2.3}$$

Dans la pratique, la forme la plus simple n'est pas efficace dans toutes les situations (ex. changement d'illumination, objets en arrière-plan instables, etc.) et plusieurs variantes sont utilisées pour identifier les objets en mouvement dans les vidéos [41].

2.3.1.2 Détection avec modélisation spatiale de l'arrière-plan

Les techniques dans cette catégorie créent un modèle d'arrière-plan stable des trames de la vidéo pour détecter les mouvements. Il s'agit dans certains cas d'utiliser la moyenne ou la médiane d'un certain nombre n de trames consécutives de la vidéo. Cette technique encore appelée *Frame Difference with Memory - FDM* [6] permet d'éliminer les bruits dans l'image lors du seuillage et permet ainsi de partitionner les trames de la vidéo en régions contiguës de changements. Yuting *et al.* [52] ont utilisé cette approche pour détecter les objets en mouvement dans une vidéo dans le but de la segmenter. Les auteurs dans [52] ont proposé de faire une estimation progressive de l'arrière-plan des images en créant des régions homogènes qui représentent les objets. Dans [11, 25], l'arrière-plan est estimé en prenant la moyenne ou la médiane de plusieurs trames consécutives de la vidéo.

2.3.1.3 Détection avec modélisation statistique de l'arrière-plan

Des modèles statistiques de l'arrière-plan peuvent être créés pour la détection des objets en mouvement. Elgammal *et al.* [18] ont proposé une méthode non-paramétrique de modélisation de l'arrière-plan qui se base sur la densité de probabilité de l'intensité d'un pixel à un instant t. Ils considèrent une historique récente des valeurs des intensités d'un pixel sur quelques trames précédentes : $I_1(X), I_2(X), \ldots, I_N(X)$. La probabilité que ce pixel ait à l'instant t une intensité $I_t(X)$ est estimée par un noyau Gaussien K de la forme :

$$P(I_t(X)) = \frac{1}{N}\sum_{i=1}^{N} K(I_t(X) - I_i(X)), \tag{2.4}$$

dans laquelle K peut suivre une loi Normale $\mathcal{N}(0, \varepsilon)$ et où ε représente la taille du noyau. Si les trois canaux de couleur sont indépendants, avec différentes tailles du noyau σ_j^2, cette densité de probabilité devient

$$P(I_t(X)) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{3}\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\frac{-(I_{t,j}(X) - I_{i,j}(X))}{2\sigma_j^2}. \tag{2.5}$$

Un pixel fait partie du modèle d'arrière-plan si sa densité de probabilité est inférieure à un seuil. Ainsi, les changements dûs par exemple au feuillage, à l'accroissement ou la diminution de l'illumination sont absorbés dans le modèle pour laisser uniquement les changements dûs au mouvement des objets. Le modèle d'arrière-plan est régulièrement mise à jour pour s'adapter aux changements. Stauffer *et al.* [50] ont également procédé par une estimation d'un modèle d'arrière-plan à l'aide de Gaussiennes pour la détection des objets en mouvement dans les vidéos. Chaque nouvelle valeur de pixel met à jour le

modèle de l'arrière-plan en fonction des paramètres des Gaussiennes, pour détecter les objets en mouvement et segmenter les trames de la vidéo. Des modèles de mouvement sont déterminés en utilisant les objets segmentés dans le but de suivre leur parcours dans la vidéo.

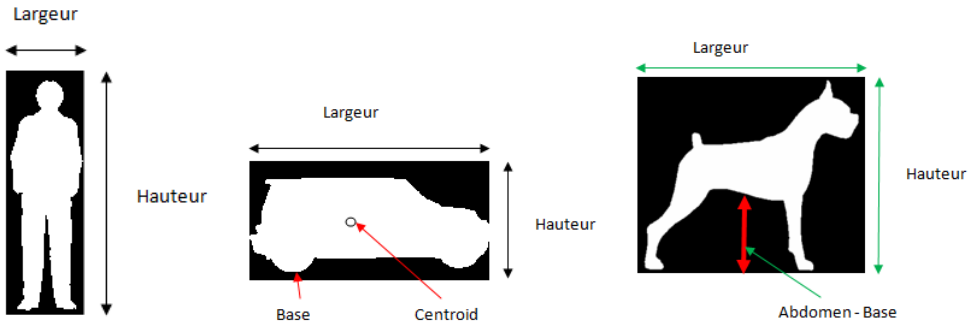
2.3.1.4 Reconnaissance des objets détectés

Le processus de reconnaissance d'objets dans les vidéos peut exploiter l'information du mouvement des objets en plus de l'information spatiale (ex., forme, silhouette) utilisée dans les images. Plusieurs travaux de recherche ont été proposés dans le passé pour la reconnaissance des objets en se basant sur l'information géométrique des objets. Dans [28], par exemple, les auteurs ont travaillé sur les propriétés géométriques des silhouettes d'objets pour identifier les humains. Ils ont suivi une approche basée sur les propriétés géométriques d'objets appartenant à différentes classes (humains, voitures et chiens). Leur travail a pour fondement les observations suivantes :

1. la largeur d'une silhouette humaine est inférieure à sa hauteur. Par contre, cette propriété est inversée pour les chiens et les voitures ;
2. le contour inférieur des silhouettes des voitures est plus bas que celui des chiens ;
3. la partie supérieure d'une silhouette humaine (en particulier la tête) a une forme unique en comparaison avec celle des voitures et des chiens.

À travers ces observations géométriques, illustrées sur les figures 2.5(a), 2.5(b) et 2.5(c), des descripteurs d'objets ont été créés en fonction des mesures suivantes : 1) le ratio largeur/hauteur pour tous les trois types d'objets, 2) la hauteur entre le contour inférieur d'un objet et le sol (pour les chiens et les voitures), 3) la distance entre le centre de gravité de l'objet et son contour inférieur (pour les chiens et les voitures) et

4) la longueur de la partie supérieure d'une silhouette humaine (représentée par le quart de sa hauteur). Dans ce travail, les tests effectués montrent que les changements de pose des silhouettes influent de façon négative sur les résultats de reconnaissance des objets.



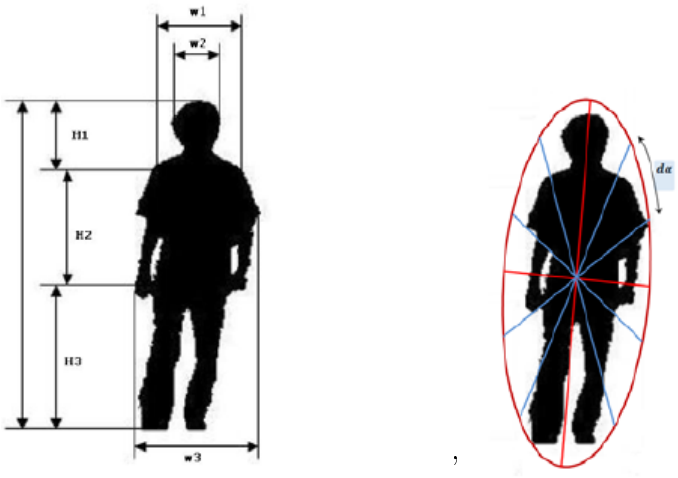
(a) Silhouette humaine (b) Silhouette de voiture (c) Silhouette de chien

FIGURE 2.5 – Différentes mesures pour la création de ratios géométriques [28].

Une autre approche de reconnaissance d'objets dans les vidéos a été proposée par Lee *et al.* [31]. Grâce à la méthode de soustraction de fond, en combinaison avec les opérations morphologiques (ex., érosion, dilatation), les silhouettes d'objets en mouvement sont isolées et identifiées selon deux classes : 1) la classe des humains et 2) la classe d'autres objets. Le descripteur ART (*Angular Radial Transformation*) [43] a été utilisé pour représenter les objets. Pedrocca *et al.* [40] ont utilisé une approche par description géométrique pour la reconnaissance des silhouettes humaines dans des vidéos. Les descripteurs d'objets sont extraits avec trois techniques différentes :

1. L'histogramme des gradients orientés appliqué à la partie supérieure des silhouettes humaines ;

2. Les ratios entre différentes parties de la silhouette humaine ; comme illustré sur la figure 2.6(a).
3. Les distances radiales normalisées entre le centre de gravité et les points sur le contour des objets.



(a) Silhouette humaine - mesures pour la création de ratios

(b) Distance radiale de la silhouette humaine

FIGURE 2.6 – Caractéristiques géométriques des humains [40].

2.3.2 Reconnaissance sans détection d'objets en mouvement

L'extraction des objets en mouvement peut se faire automatiquement en utilisant les mêmes méthodes de segmentation que dans les images. Pour guider le processus de segmentation, il est possible d'y inclure une connaissance a priori sémantique. Les techniques de segmentation semi-automatiques ont recours à des opérateurs humains pour un marquage des objets de la vidéo en fonction de leurs sémantiques. Ce procédé est bien adapté aux scènes qui sont enregistrées dans les environnements non contrôlés.

Ainsi dans les travaux [9, 8] la technique de segmentation semi-automatique a été utilisée comme préalable à la reconnaissance des objets dans des séquences de vidéos.

Parmi les techniques les plus utilisées on peut citer la reconnaissance par les points d'intérêt. Dans [8] des régions particulières ont été marquées et des points d'intérêt ont été utilisés pour suivre le parcours des objets dans la vidéo. Dans [9] cette même technique a été réalisée à partir des mouvements de la caméra utilisée pour créer les séquences de vidéo. Ces mouvements décrivent les objets dans le monde réel en générant des nuages de points en 3 dimensions (3D). Des techniques de projections des nuages de points générés ont été utilisées pour reconstruire des objets en 2 dimensions (2D) dans le but de les reconnaître. Les tests effectuées dans ce travail ont surtout montrés l'efficacité de la segmentation semi-automatique dans la reconnaissance des objets dans les vidéos, en utilisant presque exclusivement les nuages de points en 3D.

2.3.2.1 Reconnaissance par les points d'intérêt

Les points d'intérêt dans une image sont caractéristiques des discontinuités des valeurs d'intensité des pixels dans des endroits spécifiques. Différentes méthodes permettent de détecter les points d'intérêt ou les régions d'intérêt dans les images [36]. Dans la littérature on distingue trois types de détecteurs de points d'intérêt [46] : ceux qui sont basée sur les contours d'objets, ceux qui utilisent les intensités des pixels et enfin ceux qui utilisent des modèles paramétriques.

1. Détecteurs basés sur les contours : ces détecteurs passent par une première étape de détection des contours d'objets puis utilisent des méthodes pour chercher des points de courbure maximale ou d'inflexion maximale.

2. Détecteurs à base d'intensité des pixels : cette catégorie de détecteurs calcule une mesure des variations d'intensité des pixels pour isoler les points d'intérêt. Dans la majorité des cas, les points d'intérêt sont détectés sur des images en niveau de gris.

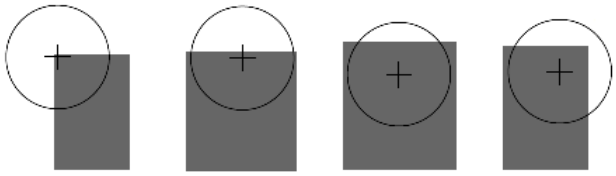
3. Détecteurs à modèles paramétriques : une image étant considérée comme un signal, cette méthode de détection utilise un modèle paramétrique du signal. Il faudra remarquer que ces détecteurs offrent une précision qui va au-delà du pixel, cependant ils sont limités à certains types de points d'intérêt [46].

Parmi les méthodes de détection de points d'intérêt les plus utilisées, on peut citer *SUZAN (Smallest Univalue Segment Assimilating Nucleus)* [48]. Cette méthode détecte les points d'intérêt en délimitant autour de chaque pixel de l'image une région circulaire ou masque. Le pixel central de cette région est considéré comme le *nucleus* et son intensité est utilisée comme *intensité référence*. Les pixels faisant partie du masque seront regroupés en deux régions selon que leur intensité est similaire ou différente de celle du *nucleus*. L'équation 2.6 est utilisée pour déterminer cette similarité. Dans cette équation, r_0 représente le *nucleus*, r représente tout point à l'intérieur du masque et t est un seuil des variations d'intensité. Le points d'intérêt sont obtenus si le nombre de points similaires au *nucleus* (calculé par l'équation 2.7) satisfait les conditions d'un minimum local et relativement à un certain seuil g défini par l'équation 2.8. Dans cette équation $u(x)$ représente le point d'intérêt. Les régions du masque ayant la même intensité que celle du *nucleus* sont appelées *UZAN (Univalue Segment Assimilating Nucleus)* et contiennent des informations sur la structure de ces régions de l'image. Les figures 2.7(a) et 2.7(b) montrent des exemples de masque et de coins *USAN*.

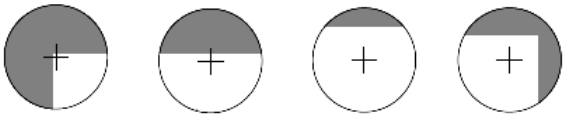
$$c(r, r_0) = e^{-(\frac{I(r)-I(r_0)}{t})^6} \tag{2.6}$$

$$n(r_0) = \sum_r c(r, r_0) \tag{2.7}$$

$$u(x) = \begin{cases} n(r_0) = g - n(r_0) & \text{si } n(r_0) < g \\ 0 & \text{sinon} \end{cases} \tag{2.8}$$



(a) Masques de SUSAN sur des images gris. Les *nucleus* sont représentés par le signe +.



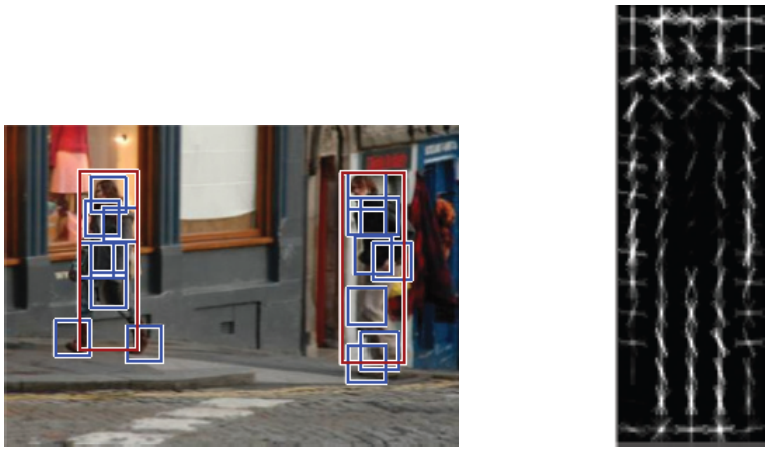
(b) Régions UZAN. Les zones en blanc ont la même intensité que le *nucleus*.

FIGURE 2.7 – Composantes du détecteur de points d'intérêt *SUZAN* [48].

2.3.2.2 Reconnaissance par l'histogramme des gradients orientés

L'histogramme des gradients orientés (HoG) calcule la distribution des gradients d'intensité sur des cellules régulières de l'image. Ces cellules font partie des fenêtres de détection qui sont délimitées sur l'image. L'orientation des gradients dans chaque cel-

lule est utilisée pour construire un histogramme de valeurs. Ces derniers constituent les descripteurs des objets détectés dans l'image. La reconnaissance d'objets dans les vidéos par utilisation du HoG a été effectuée dans [8]. Une base de données *vérité terrain* a été créée avec 32 classes d'objets différentes afin de prendre en compte une sémantique assez variée. L'histogramme des gradients orientés a également été utilisée dans [19] pour la détection des silhouettes humaines dans les images. Dans cette méthode, la représentation d'objets sous forme de modèles déformables a été utilisée. Les différentes parties des objets détectés dans les images forment les composantes des modèles. La création du modèle déformable tient compte de l'agencement spatial de ses composantes. Des fenêtres de détections de différentes tailles sont utilisées pour la création des descripteurs d'objets. La figure 2.8(a) montre deux silhouettes humaines avec des fenêtres de détection de différentes tailles. La figures 2.8(b) est une illustration des gradients ainsi que leurs orientations pour une silhouette humaine.



(a) (b)

FIGURE 2.8 – Exemple de fenêtres de détection et orientations des gradients [19].

a) Fenêtres de détection, b) gradients et leurs orientation.

2.4 Conclusion

Les méthodes de détection de changement par différence spatio-temporelle sont efficaces pour certaines catégories de vidéos non complexes, dans lesquelles on n'observe pas de mouvements significatifs d'objets dans l'arrière-plan des scènes. Elles restent cependant limitées par les problèmes d'occlusion et sont très sensibles aux bruits. Les techniques de modélisation statistique de l'arrière-plan sont plus adaptées aux vidéos ayant des scènes plus complexes, dans lesquelles on peut observer deux types de mouvement : les mouvements d'objets à identifier ainsi que des mouvements d'objets de l'arrière-plan. La modélisation statistique de Elgammal *et al.* [18], par exemple, permet de gérer les mouvements d'objets de l'arrière-plan. Cependant, cette technique ainsi que celle proposée par Stauffer *et al.* [50], ne tiennent compte que des mouvements de courte durée dans l'arrière-plan des scènes. En outre, le temps de traitement des méthodes de modélisation statistique d'arrière-plan est relativement élevé.

Certains descripteurs d'objets présentés dans ce chapitre pour la reconnaissance d'objets utilisent plusieurs ratios pour essayer de représenter avec un certain degré de fiabilité la silhouette des objets. Les descripteurs basés sur les ratios de distance sont très sensibles aux déformations observées dans la silhouette des objets. Ainsi, leur tolérance aux propriétés d'invariance (ex., changement d'échelle) est faible ; cela se justifie par le fait que les différentes mesures utilisées sont relatives et variables pour chaque classe d'objet. Ces descripteurs sont donc peu généralisables aux autres types d'objets. D'autres descripteurs utilisent par contre une approche par *patches* ou sont directement créés à partir des images en niveau de gris (exemple du *HoG* [14]). L'utilisation des images en niveau de gris entraîne cependant une plus grande sensibilité aux bruits ; ce qui n'est

pas le cas lorsque les descripteurs d'objets sont créés à partir des silhouettes extraites avec les méthodes de segmentation d'image ou de segmentation du mouvement dans les vidéos. Dans ces situations les bruits dont éliminés par le processus de segmentation.

La reconnaissance ou l'identification d'objet se base principalement sur les différents descripteurs utilisés. L'un des facteurs importants dans un processus de reconnaissance d'objet concerne le temps de traitement des algorithmes. De ce fait, la création de descripteurs à partir des images en niveau de gris et par parcours des régions d'images avec des *patches* entraînent un temps de traitement très élevé, comparativement à la création des descripteurs en fonction des silhouettes d'objets. En effet, l'utilisation des silhouettes d'objets appartenant aux images binaires (qui sont par exemple extraites de vidéos) présente l'avantage d'être plus rapide en temps de traitement, car les erreurs liées aux bruits par exemple sont déjà éliminées par la segmentation. La reconnaissance des objets dépend fortement de la technique de description choisie, qui doit pouvoir être généralisable à des objets appartenant à plusieurs classes. Elle doit également répondre à un impératif de minimiser la quantité d'informations nécessaires pour décrire l'ensemble des objets appartenant à une classe.

La majorité des travaux antérieurs cités dans ce chapitre essaient d'identifier les objets sans tenir compte de leurs changements de poses. Nous citerons néanmoins le travail de Ku *et al.* [28] qui ont essayé d'identifier les objets en fonction de leurs poses. L'information sur les différentes poses des objets peut être exploitée pour étudier l'impact des déformations des silhouettes sur la fiabilité des méthodes de description choisies.

Chapitre 3

Méthodologie

Dans ce chapitre, nous énoncerons en premier lieu les étapes qui constituent notre processus de reconnaissance d'objets en mouvement dans les vidéos. Nous présenterons ensuite en détail les éléments de notre méthodologie.

3.1 Rappel sur la problématique

Dans le chapitre précédent, nous avons présenté des travaux antérieurs sur les méthodes de détection et de reconnaissance des objets dans la vidéo. Le processus de reconnaissance des objets dans le cas de la vision artificielle est beaucoup plus complexe comparativement au système visuel humain (SVH) ; cela est dû au fait que les ordinateurs ne disposent pas des mêmes mécanismes de mémoire, d'apprentissage et de traitement que le SVH. Pour une reconnaissance efficace d'objets, il faudra créer des descripteurs adéquats.

Différents types de descripteurs ont été proposés par le passé (création de ratios, l'histogramme des gradients orientés, etc.) et permettent soit de décrire la silhouette entière de l'objet (descripteur global) ou certaines régions spécifiques (descripteur local). Le but recherché par ces travaux est la création d'un descripteur efficace pour la reconnaissance des objets connus. Il s'agit de l'un des problèmes traités dans ce travail. Cependant plusieurs limitations existent dans les travaux antérieurs. Les descripteurs

basés sur les ratios de distance, par exemple, sont très sensibles aux déformations visuelles observées dans la silhouette des objets. Ceux qui utilisent les *patches* ou qui sont créés à partir des images en niveau de gris sont sensibles aux bruits. Les descripteurs d'objets devraient être invariants aux transformations affines, telles que les changements d'échelle, de translation ou de rotation. Ils doivent également être invariants aux changements dans la pose des objets. De plus, il faudra relever les problèmes qui sont liés aux dimensions élevées des descripteurs dans les processus de reconnaissance d'objets. La dimension des descripteurs est l'un des paramètres importants pour l'efficacité de la reconnaissance d'objets par les divers algorithmes. Elle influe sur le temps de traitement des algorithmes de reconnaissance d'objets. Dans le cadre de la vidéo surveillance par exemple, les algorithmes doivent pouvoir reconnaître les objets en temps-réel.

3.2 Méthode de détection et de reconnaissance d'objets

Nous proposons un processus de détection et de reconnaissance d'objets comportant trois (3) étapes principales : pré-traitement, création de descripteurs (descripteur global et descripteurs locaux) et reconnaissance des objets. Le schéma général de notre processus est illustré sur la figure 3.1.

La première étape du processus consiste à analyser une vidéo en vue de détecter la présence d'objets en mouvement dans les scènes qui la composent et de pouvoir les isoler de l'arrière-plan. Nous aborderons dans notre méthodologie différentes techniques permettant d'isoler les objets en mouvement dans les vidéos. Les techniques utilisées

durant cette première étape dépendent de la complexité de la vidéo ; cette complexité est liée aux conditions d'enregistrement des vidéos, dont l'utilisation de cameras stables.

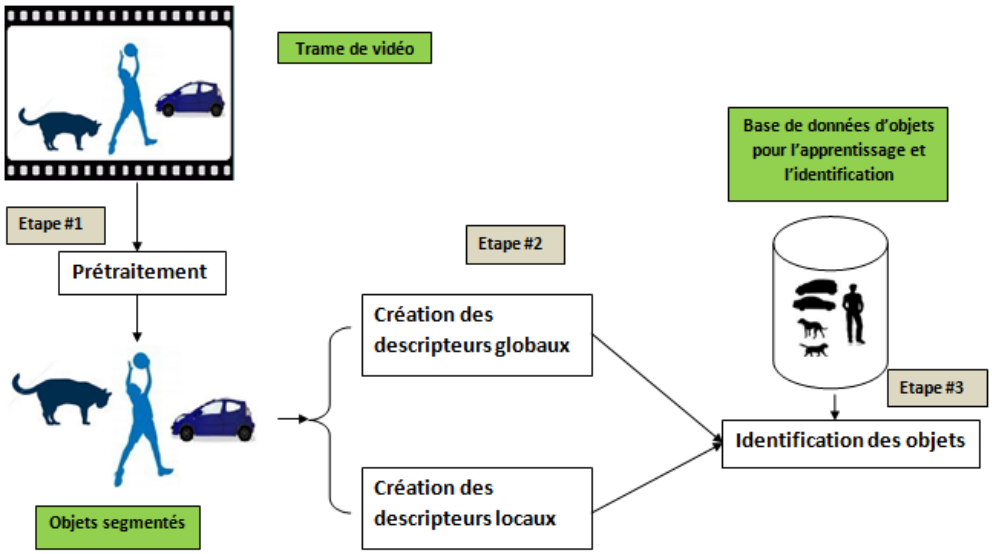


FIGURE 3.1 – Schéma général du processus.

Il faudra ainsi distinguer :

1. Les vidéos enregistrées dans des environnements contrôlés (ex., laboratoire) : dans ce genre d'environnement, il est possible d'avoir un contrôle sur l'illumination de la scène de la vidéo. L'illumination de la scène, qui définit également l'intensité des pixels de la vidéo, peut alors être considérée comme constante. Un autre aspect qui peut caractériser cet environnement est la stabilité des objets qui définissent l'arrière-plan de la vidéo.
2. Les vidéos enregistrées dans des environnements non contrôlés (ex., trafic, milieu urbain) : contrairement à l'illumination constante observée dans les environnements contrôlés, dans les conditions réelles, différents facteurs peuvent entraîner

des changements soudains ou à long terme dans l'illumination de la scène. Ces changements peuvent provenir de l'apparition de nuages qui assombrissent le paysage ou de la levée soudaine de vents qui fait bouger les feuilles et branches des arbres. Les mouvements ainsi observés rendent l'arrière-plan de la scène instable et posent des difficultés supplémentaires pour détecter les vrais objets en mouvement.

Durant la seconde étape de notre processus, nous allons représenter les objets par leurs descripteurs. Ces descripteurs ont pour but de représenter de façon unique et fiable les différents types d'objets. Un descripteur de forme d'objet est souvent évalué par le degré de fiabilité avec lequel il représente les objets. La forme ou la silhouette des objets est une caractéristique visuelle importante des objets et les descripteurs d'objets permettent de reproduire plus ou moins efficacement cette caractéristique visuelle. Différentes méthodes de description des objets existent, et doivent être indépendantes de l'application sous-jacente. Un autre facteur important dont il faudra tenir compte est le degré de complexité algorithmique nécessaire pour la description des objets.

Nous proposons une approche de reconnaissance d'objet par utilisation de deux types de descripteurs : descripteur global et descripteur local. Le descripteur global d'objet permettra de représenter la forme géométrique visuelle de l'objet. Les différentes déformations qui sont observées dans la silhouette de l'objet seront capturées dans l'information contenue dans le descripteur global. Le descripteur local d'objet permettra par contre, d'acquérir l'information sur des régions particulières des silhouettes d'objets. L'information représentée par les descripteurs locaux est caractéristique des changements qui seront observées dans des zones limitées de la silhouette des objets.

Nous utiliserons dans la troisième étape l'information contenue dans les descripteurs d'objets dans le but de les reconnaître. Cette reconnaissance s'effectuera par rapport

à une base de données d'objets connus. La reconnaissance à partir de la silhouette des objets permettra de retrouver dans la base de données d'objets, ceux dont les silhouettes sont similaires à des objets inconnus. Cette similarité entre silhouettes d'objets, relative à la perception visuelle de ces objets par les humains, devrait pouvoir être obtenue lorsqu'on est en présence d'une translation, d'une rotation ou d'une variation de l'échelle de l'objet. La reconnaissance des objets peut s'effectuer selon deux approches : la reconnaissance d'instance (ex., reconnaissance faciale) et la reconnaissance de classes [53]. La reconnaissance d'instance comme son nom l'indique permet de reconnaître un objet connu, et qui peut potentiellement être observé sous des angles différents ou en présence d'occlusion. Ce type de reconnaissance d'objet est utilisé en général pour la biométrie. La seconde méthode de reconnaissance (reconnaissance de classes) permet de reconnaître un objet comme appartenant à une classe d'objets connus (ex., Voitures, Humains, Chats). Notre travail porte sur la reconnaissance de classes.

3.3 Détection d'objets en mouvement

3.3.1 Représentation d'une vidéo

Une vidéo peut être considérée comme une séquence d'images dont chacune contient une vue statique des scènes qui s'y déroulent. Ces images, ou trames, sont prises dans des intervalles réguliers et très courts (30 images par seconde). Les trames sont des images à deux dimensions $X = (x, y)$ et l'axe temporel t constitue la troisième dimension. La figure 3.2 est une illustration de la représentation compacte d'une vidéo avec une superposition des trames qui la composent.

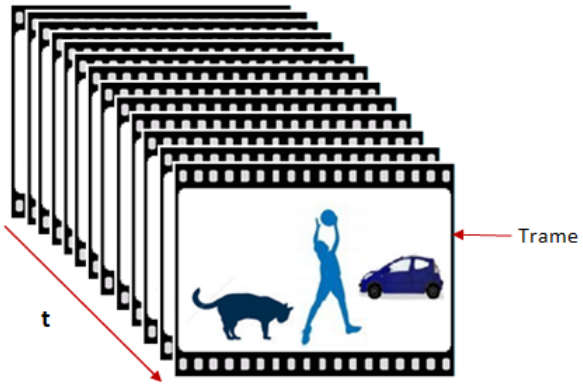


FIGURE 3.2 – Représentation compacte d'une vidéo. t représente l'axe temporel.

Chaque trame de cette vidéo peut donc être utilisée dans les opérations de pré-traitement, comme illustrée sur la figure 3.2 du schéma général du processus de reconnaissance. Ces opérations de pré-traitement ont pour but de détecter les objets en mouvement dans la vidéo.

La détection d'objets en mouvement dans la vidéo utilise des méthodes de segmentation du mouvement dans le but de marquer les régions des trames correspondant aux objets en mouvement. Nous présentons dans la section suivante les méthodes de segmentation du mouvement qui sont utilisées dans ce travail.

3.3.2 Modélisation spatiale de l'arrière-plan

La méthode *FDM - Frame Difference with Memory* permet la détection d'objets en mouvement dans les vidéos avec modélisation de l'arrière-plan de la scène de vidéo. Pour chaque trame à l'instant t de la vidéo, représentée par $I_t(X)$, on crée un modèle stable de l'arrière-plan à partir d'un certain nombre n de trames précédentes. Cette opération est effectuée selon l'équation 3.1.

$$FDM_t(X) = I_t(X) - \overline{I_t}(X), \tag{3.1}$$

avec

$$\begin{cases} \overline{I_t}(X) = (1-\alpha)I_t(X) + \alpha\overline{I_{t-1}}(X),\ t = 1, \ldots, n \\ \overline{I_0}(X) = I_0(X), \text{et} \\ 0 < \alpha < 1 \end{cases} \tag{3.2}$$

Dans cette équation α est un coefficient de pondération pour le calcul de la moyenne pondérée des valeurs de pixels des trames considérées, $\overline{I_t}(X)$ est la moyenne des trames au temps t. Lorsque $\alpha = 1$, seules les trames précédentes sont considérées dans le calcul de la moyenne. Quand $\alpha = 0$, la trame courante est considérée au détriment des trames précédentes dans le calcul de la moyenne. Il est possible d'effectuer un compromis entre les trames précédentes et courante en affectant au coefficient α une valeur telle que $\alpha = 0.5$. L'équation stipule également que la première trame est confondue à la moyenne de départ $\overline{I_0}(X)$. On peut alors appliquer un seuillage global ou adaptatif à la fonction $FDM_t(X)$. Le seuillage global consiste à fixer une seule valeur de seuil pour toute l'image. Dans le cas du seuillage adaptatif, différents seuils sont utilisés pour les régions de l'image. Cette opération permet d'éliminer le bruit et de dégager des régions homogènes contenant l'information du mouvement.

3.3.3 Modélisation de l'arrière-plan par une Gaussienne

Cette méthode consiste à effectuer une modélisation statistique des pixels composant les trames de la vidéo en fonction de leur intensité, leur couleur ou leur texture. Dans le

cas des vidéos où l'on observe de faibles mouvements dans l'arrière-plan des scènes, les valeurs des pixels des trames peuvent être modélisées par des Gaussiennes [37].

En considérant une trame de vidéo à l'instant t, l'estimation de l'arrière-plan par cette méthode s'effectue en calculant de façon récursive la moyenne μ_t et la covariance Φ_t des intensités des pixels sur un nombre n de trames précédentes par l'équation 3.3 :

$$\begin{cases} \mu_t = \alpha I_t(X) + (1-\alpha)\mu_{t-1} \\ \Phi_t = (1-\alpha)\Phi_{t-1} + \alpha(I_t(X) - \mu_t)(I_t(X) - \mu_t)^T. \end{cases} \tag{3.3}$$

Dans cette équation, $I_t(X)$ est l'intensité courante du pixel dans l'espace de couleur , μ_{t-1} est la moyenne précédente, α est un coefficient de pondération dont les valeurs sont comprises entre $[0, 1]$. Les pixels sont classés comme faisant partie de l'arrière-plan ou d'un objet en mouvement en utilisant l'estimation de vraisemblance de l'équation 3.4.

$$l(\Psi) = -\frac{1}{2}(I_t(X) - \mu_t)^T \Phi_t^{-1}(I_t(X) - \mu_t) - \frac{1}{2}\ln|\Phi_t| - \frac{d}{2}\ln(2\pi), \tag{3.4}$$

dans laquelle Ψ représente les paramètres de l'estimateur de vraisemblance, $I_t(X)$ désigne la couleur du pixel courant et μ_t la moyenne. Dans l'espace de couleur de dimension d, $I_t(X)$ suit la loi Normale multivariée $\mathcal{N}(\mu, \Phi)$, telle que la probabilité de la couleur d'un pixel soit donnée par l'équation 3.5

$$p(I_t(X)) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Phi|^{\frac{1}{2}}} \exp(-\frac{1}{2}(I_t(X) - \mu)^T \Phi^{-1}(I_t(X) - \mu)). \tag{3.5}$$

Le maximum de vraisemblance a pour objectif de trouver une quantité $l(\Psi)$ qui maximise l'estimateur $\Psi = (\mu, \Phi)$ et ainsi la probabilité pour chaque couleur $I_t(X)$ de faire partie ou non de l'arrière-plan de la vidéo. Il se calcule plus facilement en prenant

le logarithme népérien de l'équation précédente, soit :

$$l(\Psi) = \ln[p(I_t(X))]. \tag{3.6}$$

3.3.4 Modélisation de l'arrière-plan par mélange de Gaussiennes

Cette méthode est plus adaptée à la détection d'objets en mouvement lorsque l'arrière-plan contient des instabilités qui proviennent des changements dans l'illumination de la scène ou des mouvements d'objets non stationnaires. Elle prend en compte l'historique des intensités sur chaque pixel dans le temps [37]. À chaque instant t, l'historique des intensités des pixels peut être représentée par $I_1(X), I_2(X), \ldots, I_t(X)$. Chaque pixel est modélisé par un mélange de plusieurs Gaussiennes suivant l'équation 3.7.

$$p(I(X)) = \sum_{k=1}^{N} w_{k,t}(X)\chi(I(X), \mu_{k,t}(X), \Phi_{k,t}(X)). \tag{3.7}$$

Dans cette équation, $I(X)$ est un vecteur à trois dimensions pour chacune des couleurs rouge, bleu et vert, $\mu_{k,t}(X)$ et $\Phi_{k,t}(X)$ représentent respectivement la moyenne et la matrice de covariance de la k^{ieme} Gaussienne au temps t, et N est le nombre de composantes du mélange. $w_{k,t}$ est le poids de la k^{ieme} Gaussienne tel que $0 < w_{k,t} \leq 1$, ses valeurs sont comprises dans l'intervalle $[0, 1]$.

$$\sum_{k=1}^{N} w_{k,t} = 1. \tag{3.8}$$

La densité de probabilité $\chi(I(X), \mu_{k,t}(X), \Phi_{k,t}(X))$ suit une loi Normale multivariée et s'obtient par l'équation 3.5 : Les valeurs du coefficient $w_{k,t}$ sont également mises à

jour dynamiquement selon l'équation 3.9.

$$w_{k,t} = (1 - \alpha)w_{k,t-1} + \alpha(M_{k,t}), \tag{3.9}$$

où $M_{k,t}$ prend la valeur 1 si le pixel se trouve dans l'intervalle de confiance d'une Gaussienne ; dans le cas contraire sa valeur est nulle ; α est le taux d'apprentissage du modèle. À la fin de cette première étape de segmentation du mouvement, l'image binaire obtenue est composée d'objets qui doivent être isolés par segmentation.

3.3.5 Extraction des régions d'intérêt

Les images binaires obtenues à la fin du processus de segmentation du mouvement contiennent des régions qui représentent des objets identifiables. Certains de ces objets peuvent être dépourvus d'une sémantique connue. Nous parlons de sémantique connue pour caractériser les objets facilement reconnaissables par la vision humaine (ex., chiens, humains, voitures). Les régions dépourvues d'intérêt doivent être ainsi éliminées à travers des méthodes d'amélioration de la qualité.

3.3.5.1 Filtrage morphologique des images binaires

Le filtrage morphologique [6] est l'une des techniques utilisées pour le pre-traitement des images binaires dans le but d'améliorer leur qualité. De façon formelle nous pouvons définir la notion de filtrage booléen en ces termes :

Considérons une image binaire échantillonnée $I_t(X)$ dans laquelle les pixels composant les objets identifiables ont pour valeur *1* et ceux de l'arrière-plan ont pour valeur *0* (voir figure 3.3). Soit une fenêtre glissante $B = (\rho_1, \rho_2, \ldots, \rho_n)$ de dimension n. On

peut appliquer une transformation sur l'image binaire par la fenêtre glissante B à travers l'équation 3.10.

$$\begin{cases} \psi_b(I_t[X]) = b((I_t(X) \star \rho_1), \ldots, (I_t(X) \star \rho_n)) \\ \text{avec } (I_t(X) \star \rho_i) = v_i, \end{cases} \tag{3.10}$$

dans laquelle $b(v_1, \ldots, v_n)$ est une fonction booléenne à n variables $(n > 0)$ telle que $b = (AND, OR, etc.)$. L'application $I_t(X) \mapsto \psi_b(I_t(X))$ correspond au filtrage booléen. En variant les éléments de la fonction booléenne b on crée différents types de filtres booléens.

La fenêtre glissante B est désignée sous le nom d'*élément structurant*. Dans le cadre du filtrage morphologique binaire, les composantes de l'*élément structurant* prennent également leurs valeurs dans l'ensemble $\{0, 1\}$.

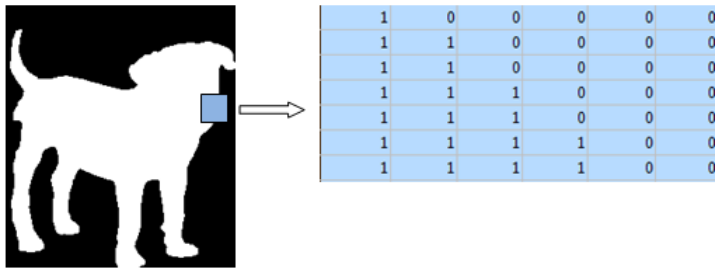


FIGURE 3.3 – Image binaire avec une silhouette de chien avec les valeurs des pixels de l'échantillon en bleu

Différents types de filtrages booléens peuvent être appliqués aux images binaires dans le but d'extraire les objets (ex. *dilatation, érosion, fermeture, clôture*). Supposons que les pixels de l'objet forment l'ensemble A et ceux de l'arrière-plan forment son complément $\overline{A}$. La dilatation de l'objet par une fenêtre B s'obtient par

$$A \oplus B \equiv \{X + Y : X \in A, Y \in B\} = \bigcup_{Y \in B} A_{+Y}, \tag{3.11}$$

où $A_{+Y} \equiv \{X + Y : X \in A\}$ est la translation de A le long du vecteur Y. De même, si $B^T \equiv \{X : -X \in B\}$ représente la réflexion de B par rapport à son origine, la transformation de A par B^T correspond à l'érosion et s'obtient par

$$A \ominus B \equiv \{X : B_{+X} \subseteq A\}. \tag{3.12}$$

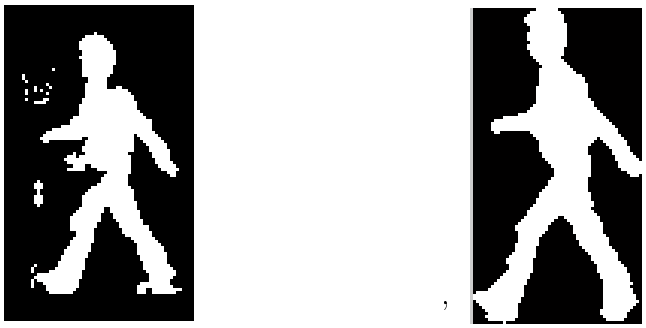
Les combinaisons de l'érosion et de la dilatation créent deux autres opérations. La première, l'*ouverture*, est définie par $A \circ B \equiv (A \ominus B) \oplus B$; la seconde, la *clôture* s'obtient par $A \bullet B \equiv (A \oplus B) \ominus B$ [6].

Les filtres morphologiques binaires agissent sur la forme des objets. Le processus de dilatation permet de lisser le contour des objets, et supprime également les trous de certaines tailles qui sont présents à l'intérieur des objets. L'érosion effectue également le lissage des contours en supprimant les légères bifurcations ou prolongements inutiles.

3.3.5.2 Sélection des composantes connectées

Il est souvent désirable de ne conserver que les objets dans l'image binaire ayant une certaine taille S. Dans cette situation, un algorithme simple consisterait à supprimer les objets dans l'image dont la taille serait inférieure à S. La sélection des composants connectés permet de supprimer les objets dépourvus de sémantique, caractérisant la présence de bruits par exemple et donc non identifiables.

La figure 3.4 montre un exemple de filtrage morphologique d'image binaire. Sur la figure 3.4(a) on peut apercevoir différentes régions dont l'une correspond à une silhouette humaine.



(a) Image avant filtrage

(b) Image après filtrage

FIGURE 3.4 – Exemple de filtrage morphologique sur une image binaire.

La seconde partie de notre travail consiste à utiliser des techniques appropriées pour décrire les objets. Les informations nécessaires pour identifier les objets en mouvement seront obtenues par l'utilisation de méthodes de description de leur contours ou de régions particulières de ces objets ; ces méthodes font l'objet de la section suivante. Les descripteurs permettent de représenter les objets extraits des trames de la vidéo, sous une forme facile à manipuler par les algorithmes de reconnaissance. Deux types de descripteurs sont considérés dans la reconnaissance d'images et des objets : les descripteurs globaux et les descripteurs locaux.

3.4 Descripteur global d'objets

3.4.1 Introduction

Un descripteur global d'objet utilise la représentation de la silhouette entière des objets pour les décrire. Ce type de descripteur donne ainsi une indication sur la forme géométrique globale de l'objet (ex. *contour*). D'une manière générale, les descripteurs globaux (ou descripteurs de formes) doivent posséder les caractéristiques suivantes :

- permettre d'établir une représentation plus ou moins fidèle de la forme ou du contour de l'objet ;
- permettre d'obtenir une discrimination des différentes formes d'objets (ex. *humains, chiens*) ;
- posséder des propriétés d'invariance au changement d'échelle, à la translation et à la rotation.

Le descripteur global retenu dans ce travail est la *forme contextuelle* [4].

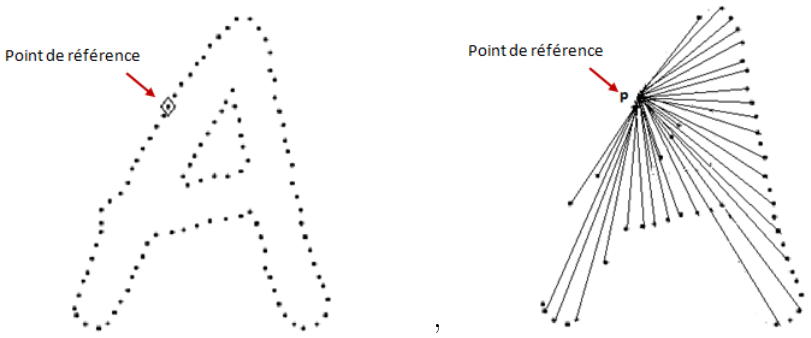
3.4.2 La forme contextuelle

3.4.2.1 Modèle théorique et création

Les objets extraits du processus de prétraitement peuvent être représentés par l'ensemble N des points sur leurs contours noté Γ). De façon formelle, cet ensemble de points finis et ordonnés p_i peut être modélisé par l'équation 3.13.

$$\Gamma = \{p_i = (x_i, y_i) \text{ avec } i = 1, \ldots, N\}\,. \tag{3.13}$$

À partir d'un point de référence $p_i = (x_i, y_i)$, on peut tracer l'ensemble des vecteurs qui le relie aux $N - 1$ points du contour. Ces vecteurs expriment la configuration du contour par rapport au point de référence. La figure 3.5(a) montre un exemple d'objet représenté par les points du contour ainsi qu'un point de référence p_i. La figure 3.5(b) montre les vecteurs reliant le point de référence aux autres points du contour. La forme contextuelle (FC) est par conséquent la distribution des points sur le contour de l'objet par rapport à chaque point de référence pris individuellement.



(a) Points sur le contour

(b) Vecteurs liant $N - 1$ points au point de référence

FIGURE 3.5 – Contour et vecteurs exprimant sa configuration dans la FC.

Ce descripteur global d'objet (FC) est obtenu par l'utilisation d'un système de coordonnées polaires ayant pour origine les coordonnées des points de référence p_i du contour. Ce système de coordonnées polaires, illustré sur la figure 3.6, est composé d'un ensemble de M cercles de rayons Δr croissants ; ces cercles sont ensuite subdivisés en N angles $\Delta\theta$ constants. Le système permet de regrouper des sous-ensembles de points sur le contour de l'objet dans un histogramme de c cases ou bin avec $c = M \times N$. Il est ainsi

possible d'obtenir différentes formes contextuelles selon le positionnement de ce système de coordonnées. De façon formelle, cet histogramme est défini par l'équation 3.14 :

$$h_i(c) = \sharp \{q \neq p_i : (q - p_i) \in bin(c) \text{ et } q \in (n - p)\}, \tag{3.14}$$

dans laquelle $h_i(c)$ représente l'histogramme des points sur le contour, $bin(c)$ représente une entrée de l'histogramme, p_i représente un point de référence et q fait partie des autres points du contour.

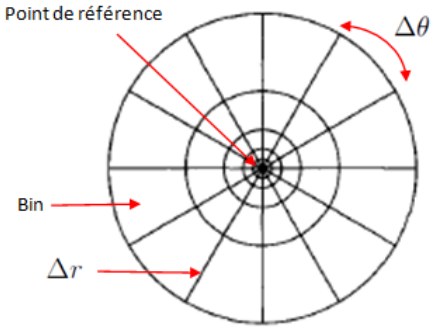


FIGURE 3.6 – Système de coordonnées dans la forme contextuelle. Δr représente un des rayons, $\Delta\theta$ indique l'angle constant.

3.4.2.2 Création de la nouvelle forme contextuelle

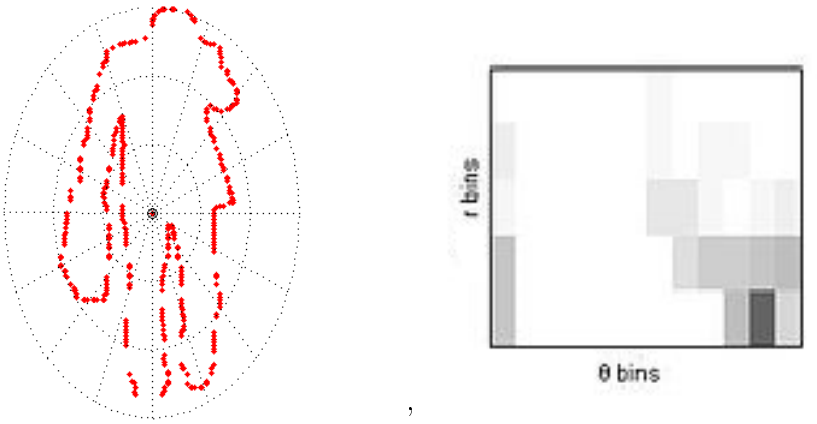
La création de ce descripteur global s'effectue par le positionnement du système de coordonnées sur le centre de gravité des objets. Le principe général utilisé dans notre méthodologie consiste à compter les points dans chaque région du système de coordonnées et à insérer le total correspondant dans les cases de l'histogramme. La

nouvelle forme contextuelle utilise un seul point de référence qui est le centre de gravité des objets.

La création des vecteurs qui expriment la configuration des points sur le contour par rapport au centre de gravité des objets a été effectuée en utilisant la distance de *Mahalanobis*. Considérons une distribution statistique de N vecteurs $U = (u_1, u_2, \ldots, u_d)$ de d dimensions, admettant une moyenne μ et une matrice de covariance Φ. Par définition, la distance de *Mahalanobis* entre un vecteur U et la moyenne μ peut s'obtenir par l'équation 3.15 :

$$d_m(U, \mu) = \sqrt{(U - \mu)^T \Phi^{-1} (U - \mu)}. \tag{3.15}$$

Cette distance est invariante à l'échelle et prend en compte la corrélation entre les distances. Dans l'équation ci-dessus, Φ représente la matrice de covariance des distances. Dans le cas où cette matrice de covariance est une matrice identité, la distance de *Mahalanobis* est équivalente à la distance Euclidienne. L'utilisation de la distance de *Mahalanobis* permet à la forme contextuelle de s'adapter à la forme des objets extraits des trames des vidéos. La figure 3.7(a) montre un exemple de création de la FC pour une silhouette humaine. La forme contextuelle (FC) obtenue peut être affichée comme illustrée sur la figure 3.7(b).



(a) Création de la forme contextuelle

(b) Affichage de la forme contextuelle de (a). *r exprime* $\sharp\Delta r$, θ *exprime* $\sharp\Delta\theta$

FIGURE 3.7 – Création et affichage de la forme contextuelle.

3.4.2.3 Caractéristiques

La FC dans notre méthodologie présente les caractéristiques suivantes :

- Invariance aux transformations affines : cette caractéristique peut être vérifiée par l'utilisation des coordonnées relatives des points du contour de l'objet par rapport à son centre de gravité. Par conséquent, tout déplacement dans le plan n'a pas d'effet sur la forme contextuelle. Dans le but d'assurer une invariance à l'échelle, nous avons normalisé chaque distance par rapport à la moyenne des distances calculées.
- Descripteur de dimension réduite : dans sa forme originale décrite dans les travaux de Belongie *et al.* [4] sous l'appellation du *shape context*, ce descripteur est calculé en utilisant η points de références du contour d'objet ($\eta > 1$). Il en découle un descripteur global de dimension $\eta \times N \times M$. La forme contextuelle proposée dans

cette méthodologie par contre utilise un seul point de référence correspondant au centre de gravité des objets. Elle a pour dimension $N \times M$ de moindre taille que l'originale.

- Un des avantages de la nouvelle forme contextuelle proposée est lié à l'influence de la dimension du descripteur au temps de traitement des algorithmes de reconnaissance. D'une manière générale, la manipulation de structure de données de taille réduite entraîne une diminution de l'espace de traitement mémoire nécessaire et également du temps de traitement requis par les algorithmes. La description d'objets par la forme contextuelle proposée est en outre généralisable, dans ce sens tout objet extrait d'une image ou d'une vidéo peut être représenté de façon plus ou moins précise en générant sa FC à partir de son centre de gravité.
- L'utilisation de la distance de *Mahalanobis* permet de créer un descripteur qui s'adapte à la forme des silhouettes d'objets (ex. des différentes grosseurs d'une silhouette humaine). En comparaison avec la distance Euclidienne, la distance de *Mahalanobis* exprime mieux la répartition des points sur le contour par rapport au centre de gravité qui représente le point de référence.

3.4.3 Reconnaissance d'objets avec la forme contextuelle

3.4.3.1 Approche générale du processus

La reconnaissance d'un objet (inconnu) en mouvement dans une vidéo s'effectue à travers un processus de calcul de distance entre son descripteur et des descripteurs d'objets connus et répertoriés dans une base de données. Dans la littérature, on parle également de *classification*, processus permettant d'affecter des objets à leurs classes

d'appartenance. Les descripteurs (ou *features* en anglais) sont des vecteurs dans des espace à d dimensions de la forme

$$U = (u_1, u_2, \ldots, u_d)^T. \tag{3.16}$$

La dimension du descripteur influe de manière significative sur le temps de traitement des algorithmes de reconnaissance et sur leur efficacité. Considérons l'ensemble $\mathcal{W}$, de dimension z fini, des classes d'objets devant être utilisées pour la reconnaissance, avec $\mathcal{W} = \{\omega_1, \omega_2, \ldots, \omega_z\}$. La reconnaissance par le calcul de distances entre descripteurs consiste à trouver un ensemble $\mathcal{L}$ de fonctions $dist_1(X), dist_2(X), \ldots, dist_{\mathcal{L}}(X)$ telles que si un descripteur U appartient à une classe ω alors l'équation 3.17 est vérifiée.

$$dist_i(U) \leq dist_j(U) \text{ avec } j = 2, 3, \ldots, \mathcal{L}; j \neq i. \tag{3.17}$$

En d'autres termes, un algorithme de reconnaissance affecte un descripteur d'objet inconnu à une classe ω si sa distance par rapport aux autres descripteurs est la plus faible. Dans ce travail, nous utiliserons les méthodes statistiques présentées dans les sections ci-dessous.

3.4.3.2 La méthode des *K Plus Proches Voisins* (KPPV)

La méthode des *KPPV* est une méthode d'apprentissage supervisée non paramétrique [16]. Contrairement à l'apprentissage non-supervisé, l'apprentissage supervisé consiste à fournir à l'algorithme de reconnaissance des échantillons préalablement étiquetés devant lui permettre de construire un modèle d'apprentissage. Ces étiquettes constituent les classes.

Considérons un échantillon de données composé de paires de valeurs (U_i, ω_i), avec $i = 1, \ldots, n$ tel que U_i représente le descripteur de rang i, et ω_i sa classe. Soit la métrique de distance $d(U_1, U_2)$ mesurant la distance entre deux valeurs quelconques U_1 et U_2. Pour chaque descripteur inconnu U, l'algorithme des *K Plus Proches Voisins* calcule une valeur $\delta_i = d(U_i, U)$ et détermine ensuite les indices $\{a_1, \ldots, a_k\}$ des k plus petites valeurs δ_i telles que :

$$\begin{cases} \delta_{a_i} \leq \delta_j \text{ pour tout } j \notin a_1, \ldots, a_k, i \in 1, \ldots, k \text{ et} \\ \delta_{a_1} \leq \delta_{a_2} \leq \cdots \leq \delta_{a_k}. \end{cases} \tag{3.18}$$

On définit k_j comme étant le nombre de descripteurs connus qui sont les plus proches voisins de U et dont la classe est ω_i par l'équation 3.19

$$k_j = \sum_{i=1}^{k} I(\omega_{ai} = j), \tag{3.19}$$

dans laquelle $I(a = b)$ est une fonction prenant la valeur 1 si $a = b$ et 0 dans le cas contraire. La règle de décision de la méthode des *KPPV* consiste alors à associer à U l'étiquette la plus dominante parmi les k_j voisins.

3.4.3.3 Choix de la valeur de k pour le KPPV

Le nombre k est en général choisi de façon à trouver un compromis entre une valeur assez grande dans le but de diminuer la sensibilité aux erreurs de reconnaissance (notamment le vote majoritaire) et une valeur assez faible permettant d'affecter le descripteur à la classe appropriée. Ainsi, l'on peut donner à k des valeurs supérieures ou égales à 1. Dans le cas où la reconnaissance s'effectue entre deux classes, des valeurs impaires de k

permettent d'éviter l'impossibilité de prise de décision. La figure 3.8 montre un exemple de classificateur binaire pour lequel $k = 3$.

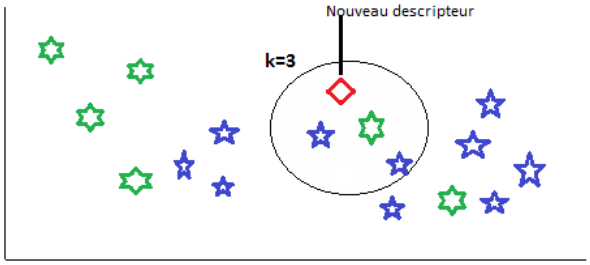


FIGURE 3.8 – Illustration de l'algorithme KPPV. L'étiquette bleu sera affectée au nouveau descripteur.

3.4.3.4 La méthode naïve de Bayes pour la classification

La classification par la méthode naïve de Bayes a pour fondement le théorème de Bayes de deux événements quelconques A et B qui s'exprime sous la forme suivante :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.20}$$

Dans cette équation $P(A|B)$ indique la probabilité a posteriori d'observer l'événement A sachant que l'événement B est réalisé. $P(A)$ et $P(B)$ sont les probabilités a priori des événements A et B.

Supposons que $p(w_k|U)$ soit la probabilité a posteriori qu'un descripteur $U = (u_1, u_2, \ldots, u_d)^T$ appartienne à une classe w_k. D'après l'équation 3.20 cette probabilité s'exprime sous la forme :

$$p(w_k|U) = \frac{p(U|w_k)p(w_k)}{p(U)}. \tag{3.21}$$

La règle de décision d'un classificateur de Bayes repose sur le maximum de la probabilité a posteriori $p(w_k|U)$. En d'autres termes, si z représente le nombre de classes, un descripteur U appartiendra à une classe w_i si :

$$p(w_i|U) > p(w_j|U) \text{ avec } 1 \leq j \leq z, \text{ et } j \neq i. \tag{3.22}$$

La particularité de la méthode naïve de Bayes est relative à l'évaluation de $p(U|w_k)$. En effet, cette méthode repose sur la forte hypothèse que les éléments d'un descripteur $U = (u_1, u_2, \ldots, u_p)^T$ sont conditionnellement indépendants, pour une classe w_k. Cette hypothèse permet ainsi d'écrire :

$$p(U|w_k) = \prod_{i=1}^{n} p(u_i|w_k). \tag{3.23}$$

La règle de décision par la méthode naïve de Bayes se définit alors par l'équation :

$$p(w_k|U) = \arg\max_{w_k} \prod_{i=1}^{n} p(u_i|w_k)p(w_k). \tag{3.24}$$

La probabilité $p(w_k)$ peut être estimée en prenant le rapport entre le nombre d'objets appartenant à la classe w_k et le nombre total d'objets. Dans ce travail, cette probabilité a été estimée selon l'équation 3.25. dans laquelle la valeur $|\mathcal{W}|$ représente le nombre de classes utilisées pour la reconnaissance des objets ; de ce fait $|\mathcal{W}| = 3$ pour les classes *Chiens, Humains et Voitures.*

$$p(w_k) = \frac{1}{|\mathcal{W}|}. \tag{3.25}$$

3.5 Descripteur local d'objet : les points d'intérêts

Comme nous l'avons mentionné dans la section 3.2 de ce chapitre, la création des descripteurs locaux d'objets intervient également pendant la seconde étape du processus de reconnaissance des objets (voir figure 3.1). Il faut noter que la plupart des applications des points d'intérêts ont pour objectifs de trouver des similarités entre des images qui peuvent être en couleur ou en niveau de gris. Dans notre cas, les descripteurs locaux sont créés dans différentes régions des images binaires obtenues à la fin du processus de pré-traitement des trames de la vidéo. Le détecteur de points d'intérêt retenu dans notre méthodologie se classe parmi ceux qui utilisent les variations d'intensité des pixels (voir section 2.3.2.1 du chapitre 2). Il s'agit du détecteur *SURF - Speeded Up Robust Features* qui est souvent privilégié par rapport aux autres détecteurs de points d'intérêt de la même catégorie, en raison de son temps de traitement assez faible [3]. Cette propriété fait du *SURF* un détecteur approprié pour les applications temps réels. Le descripteur *SURF* est invariant aux changements d'illuminations dans les images et aux transformations géométrique d'échelle, de translation et de rotation.

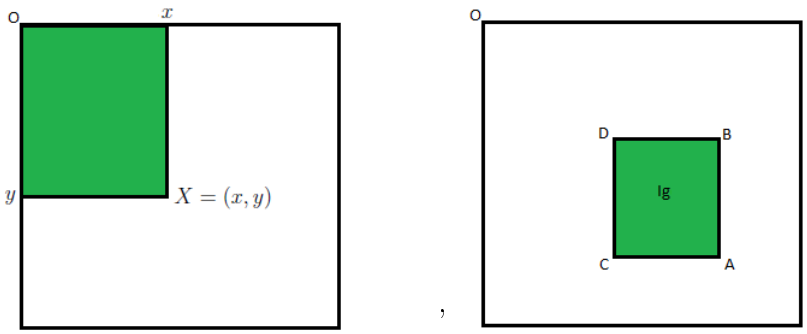
3.5.1 Modèle théorique du descripteur local

Le détecteur *SURF* utilise les techniques d'intégrale d'image et différents filtres pour localiser les points d'intérêt.

3.5.1.1 L'intégrale d'image

Une intégrale d'image $I_{integ}(X)$ en un point X de coordonnées (x, y) représente la somme des pixels d'une région rectangulaire partant de l'origine de l'image jusqu'à ce point. Elle s'obtient par l'équation 3.26 et est illustrée sur la figure 3.9(a).

$$I_{integ}(X) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(x, y). \tag{3.26}$$



(a) Intégrale au point X (b) Intégrale du rectangle

FIGURE 3.9 – Intégrales d'image.

L'intégrale d'une région rectangulaire délimitée dans une image est illustrée sur la figure 3.9(b). Elle se calcule par trois opérations élémentaires entre les intégrales d'images des quatre coins du rectangle : $I_g = A - B - C + D$, où chacun des termes A, B, C et D correspond à l'intégrale d'image définie à cette coordonnée dans l'image. Le temps de calcul d'une intégrale d'image est indépendant de la taille des rectangles qui la définissent.

3.5.1.2 Détection et description du point d'intérêt

Le SURF a pour fondement le détecteur *Hessien-Laplacien* qui utilise les matrices Hessiennes et leurs déterminants dans le but de sélectionner les points d'intérêt à différentes échelles [3]. La matrice Hessienne d'un point $X = (x, y)$ dans une image donnée, sur une échelle σ s'obtient par l'équation 3.27.

$$\mathcal{H}(X, \sigma) = \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix}. \tag{3.27}$$

Les valeurs $L_{xx}(X, \sigma)$, $L_{xy}(X, \sigma)$, $L_{yx}(X, \sigma)$, $L_{yy}(X, \sigma)$ représentent les produits de convolution Gaussiennes de second ordre $\frac{\partial^2}{\partial x^2} g(\sigma)$, $\frac{\partial^2}{\partial x \partial y} g(\sigma)$, $\frac{\partial^2}{\partial y \partial x} g(\sigma)$ et $\frac{\partial^2}{\partial y^2} g(\sigma)$ de l'image au point X. Ces valeurs sont indiquées par l' équation 3.28. σ représente l'écart-type du noyau Gaussien.

$$\begin{cases} L_{xx}(X, \sigma) = \frac{\partial^2}{\partial x^2} g(\sigma) * I, \\ L_{xy}(X, \sigma) = \frac{\partial^2}{\partial xy} g(\sigma) * I, \\ L_{yx}(X, \sigma) = \frac{\partial^2}{\partial yx} g(\sigma) * I, \\ L_{yy}(X, \sigma) = \frac{\partial^2}{\partial y^2} g(\sigma) * I. \end{cases} \tag{3.28}$$

Les valeurs de la matrice Hessienne $H(X, \sigma)$ sont approximées par des filtres rectangulaires de différentes tailles et des intégrales d'images ; ce qui permet de réduire le temps de calcul. On distingue ainsi des filtres de tailles 9×9, 15×15, 21×21, 27×27, 39×39, 51×51, etc. Le filtre de taille 9×9 correspond à l'échelle la plus faible avec un écart-type $\sigma = 1.2$. Les nouvelles valeurs de la matrice Hessienne après approximation

par les filtres sont notées D_{xx}, D_{xy} et D_{yy}. Ces valeurs correspondent à la somme des termes sur le voisinage des points X selon l'équation 3.28. Ces dernières sont ensuite utilisées pour normaliser le déterminant de la matrice Hessienne en fonction d'un facteur w représentant les différentes réponses aux filtres ; cette normalisation du déterminant a pour résultat la nouvelle valeur obtenue par l'équation 3.29.

$$\textit{déterminant approximée} = D_{xx}D_{yy} - (\delta D_{xy})^2. \tag{3.29}$$

Le facteur δ change théoriquement avec les différentes valeurs de σ, cependant il peut être maintenu constant sans aucune influence sur les résultats des approximations ; il est calculé par l'équation 3.30 [3].

$$\delta = \frac{|L_{xy}(1.2)|_F |D_{xx}(9)|_F}{|L_{xx}(1.2)|_F |D_{xy}(9)|_F} \simeq 0.9, \tag{3.30}$$

où $|\cdot|_F$ représente la norme de Frobenius. La norme de Frobenius est une norme matricielle qui peut être calculée pour une matrice $\mathcal{A}$ de dimension $m \times n$, dont les éléments sont notés sous la forme a_{ij}, par l'équation 3.31.

$$\|\mathcal{A}\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}}. \tag{3.31}$$

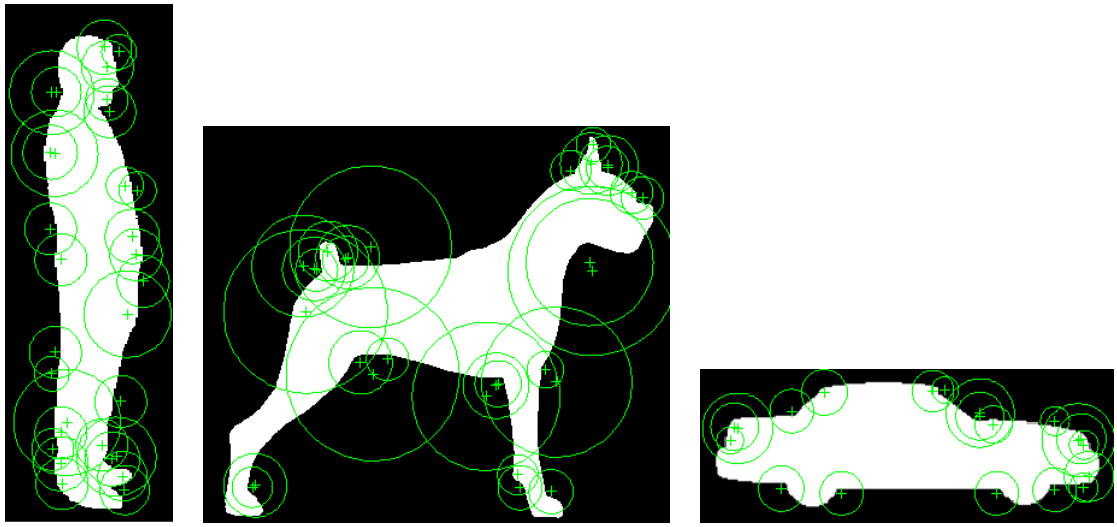
L'approximation du déterminant de la matrice Hessienne par le facteur δ devient alors celle obtenue par l'équation 3.32.

$$\textit{déterminant approximée} = D_{xx}D_{yy} - (0.9D_{xy})^2. \tag{3.32}$$

Ces nouvelles valeurs des approximations de la matrice Hessienne, qui caractérisent chaque point $X = (x, y)$ dans l'image sont enregistrées dans des structures de données appelées *blob response map*. Des opérations de suppression de maxima locaux sont ensuite appliquées aux *blob response map* pour déterminer les coordonnées et la taille des meilleurs points d'intérêt.

3.5.2 Points d'intérêts des objets en mouvement

Pour rappel, les points d'intérêts dans notre travail sont détectés dans les images binaires contenant les objets identifiables. Dans notre approche, le processus de détection des points d'intérêt est entièrement automatique. La figure 3.10 montre des exemples d'objets avec leurs points d'intérêts.

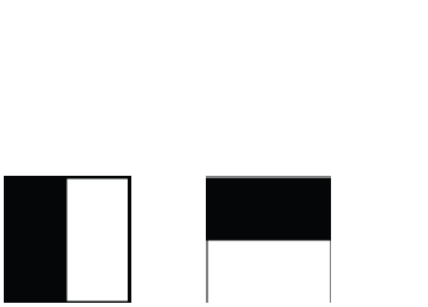


(a) Points SURF - Homme (b) Points SURF - Chien (c) Points SURF - Voiture

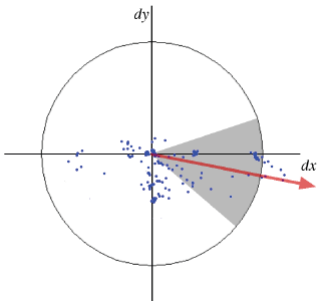
FIGURE 3.10 – Exemples de Points SURF sur des objets.

3.5.3 Création des descripteurs locaux

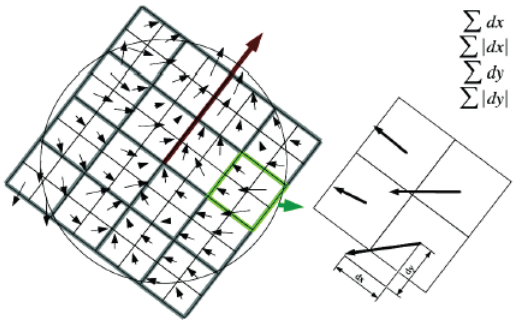
Chaque point d'intérêt *SURF* est décrit par un descripteur qui est un vecteur à 64 dimensions. Pour l'obtenir, le *SURF* détermine dans un premier temps l'orientation de régions entourant les points d'intérêt. Le procédé consiste à délimiter des régions circulaires de rayons $6s$, centrées sur les coordonnées des points d'intérêt localisés dans la section précédente ; s représentant l'échelle à laquelle le point d'intérêt a été détecté. L'orientation est obtenue par application des ondelettes de *Haar* (figure 3.11(a)) dans les directions x et y à la région circulaire, et par analyse des réponses d'un filtrage Gaussien d'écart-type $\sigma = 2.5s$.



(a) Filtres de Haar



(b) Orientation du point d'intérêt



(c) Calcul des sous-descripteurs

FIGURE 3.11 – Filtres de Haar, orientation des réponses et sous-descripteurs.

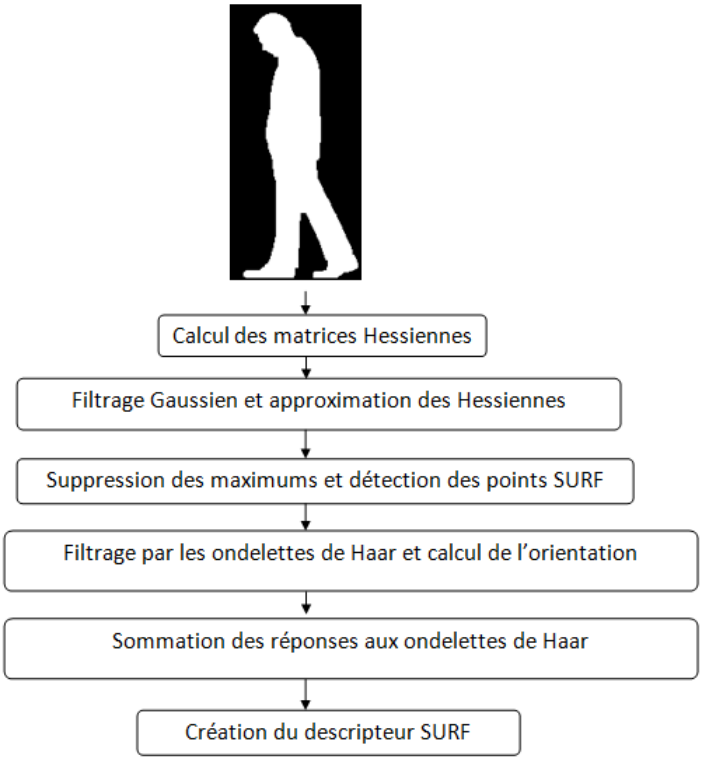


FIGURE 3.12 – Étapes de création du descripteur SURF

Les réponses du filtre sont représentées par des vecteurs dans un espace à partir desquels l'orientation dominante est déterminée (figure 3.11(b)). Pour créer le descripteur, la première étape consiste à délimiter des fenêtres carrées de taille $20s$ sur les points d'intérêt et à les aligner selon l'orientation précédente. Ces fenêtres sont subdivisées en 16 sous-régions dont chacune est divisée en quatre cellules. Dans chaque sous-région on détermine deux valeurs d_x et d_y qui correspondent respectivement aux réponses des ondelettes de *Haar* suivant la direction horizontale x, et verticale y. Ces réponses sont pondérées par une Gaussienne puis sommées pour chaque sous-région afin d'obtenir des sous-descripteurs $v = (\sum_{d_x}, \sum_{d_y}, |\sum_{d_x}|, |\sum_{d_y}|)$. Les valeurs absolues permettent de

tenir compte de la polarité des changements d'intensité des pixels. Le descripteur *SURF* à 64 dimensions est formé par l'ensemble des sous-descripteurs v qui sont obtenus pour les 16 sous-régions. L'orientation dominante ainsi que les 16 sous-régions utilisées pour le calcul de ces sous-descripteurs sont illustrées sur la figure 3.11(c). Les étapes du processus de création et de description des points d'intérêt *SURF* sont illustrées sur la figure 3.12.

3.5.4 Reconnaissance d'objets avec les points d'intérêt

Différentes méthodes existent pour la reconnaissance des objets par leurs points d'intérêt. Dans la littérature, on distingue entre autres les méthodes d'appariement ou de mise en correspondance de points d'intérêt ainsi que la technique des *Bag of Features (BoF)*.

3.5.4.1 Reconnaissance par la mise en correspondance

Cette méthode de reconnaissance d'objet a pour principe la mise en correspondance des descripteurs *SURF* d'un objet inconnu avec les descripteurs *SURF* d'objets répertoriés dans notre base de données. Cette mise en correspondance de points d'intérêts détectés dans deux images permet de mesurer le degré de similarité entre ces images. Pour un point d'intérêt quelconque P_1 pris dans une image $I_1(X)$, il s'agira de déterminer le point homologue dans une image $I_2(X)$ ou dans un ensemble d'images $I_2(X), I_3(X), \ldots, I_n(X)$ présentes dans une base de données. Ce principe est illustré sur la figure 3.13 entre deux objets.

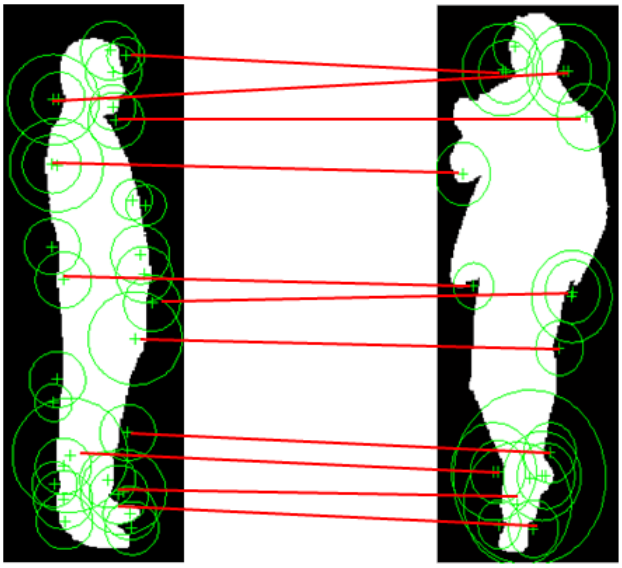


FIGURE 3.13 – Mise en correspondance de points SURF.

La mise en correspondance engendre des valeurs de similarité (*scores*) entre les descripteurs *SURF*. Ces scores de similarité entre descripteurs *SURF* sont obtenus par différentes métriques de corrélation entre les valeurs des descripteurs. Les différents algorithmes se baseront ainsi sur le minimum des scores de similarités générés par ces métriques de distance pour identifier les objets ; les scores élevés traduisent une grande dissimilarité entre les objets.

3.5.4.2 Problématique de la mise en correspondance

La mise en correspondance des points SURF s'effectue, comme indiquée précédemment, par des métriques de distance ou de corrélation [23] entre les descripteurs. Les métriques de distance suivantes sont les plus couramment utilisées :

– **Somme des carrées des différences (ou Sum of Squared Difference - SSD)**
Cette fonction mesure la similarité entre deux descripteurs par le calcul du carrée de la distance entre les valeurs de leurs attributs selon l'équation 3.33

$$SSD(d_1, d_2) = \sum |d_1 - d_2|^2. \tag{3.33}$$

– **Corrélation croisée centrée (ou Normalized Cross Correlation - NCC)**
Elle mesure la similarité entre deux descripteurs $SURF$ par leur produit scalaire selon l'équation 3.34. Les valeurs sont comprises dans l'intervalle $[-1, 1]$. Une correspondance parfaite est obtenue lorsque le résultat de la corrélation croisée centrée vaut 1.

$$NCC(d_1, d_2) = \frac{d_1.d_2}{|d_1||d_2|}. \tag{3.34}$$

Dans certaines situations, ces métriques peuvent engendrer des erreurs de reconnaissance. En effet, un descripteur $SURF$ situé au niveau de la partie supérieure d'une silhouette humaine peut correspondre à un autre descripteur $SURF$ qui est par contre localisé au niveau des pattes d'un chien. Nous illustrons cette situation par la figure 3.14 ci-dessous.

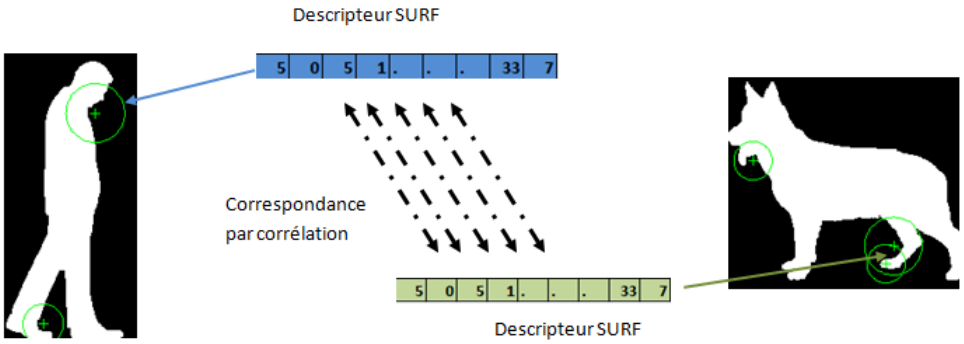


FIGURE 3.14 – Erreur possible - Corrélation des descripteurs SURF.

Sur cette figure, on remarque qu'une corrélation entre le descripteur *SURF* de la tête humaine et celui du pied du chien peut conduire à une correspondance parfaite. Cela amènera le classificateur à considérer les deux objets comme identiques, conduisant ainsi à une erreur de reconnaissance.

3.5.4.3 Mise en correspondance par groupes de points d'intérêt

Pour résoudre ce problème d'erreurs potentielles de reconnaissance, nous avons introduit une méthode de division virtuelle des objets en blocs. Cette méthode nous permet d'utiliser la distribution des points d'intérêt par blocs et d'effectuer la mise en correspondance par agrégation des scores de similarité. Notre méthode permet également de créer des scores géométriques que nous utiliserons pour pondérer les scores de similarité entre descripteurs. La figure 3.15 donne une illustration de cette division virtuelle et montre une répartition des points *SURF* dans 8 blocs différents.

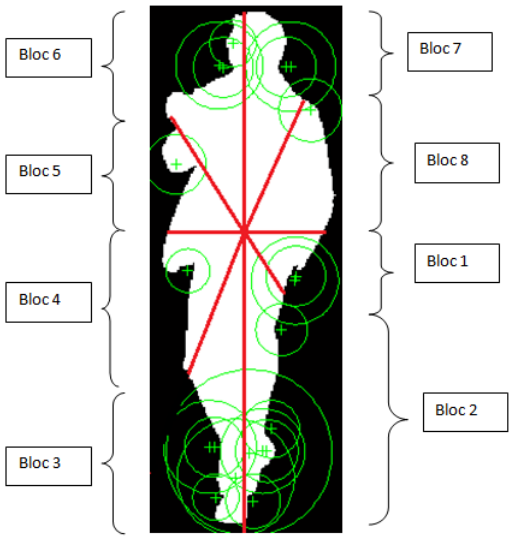


FIGURE 3.15 – Répartition des points SURF par blocs.

Chaque objet est divisé en un nombre $\mathcal{B}$ de blocs, tels que $\mathcal{B} = \{\beta_1, \beta_1, \ldots, \beta_k\}$. Pour un bloc β_r issu de la subdivision de l'image test, nous réalisons la correspondance avec ses homologues β_t d'objets dans la base de données. Soit ψ_i le score de similarité entre deux blocs homologues de deux images. Nous désignons également par τ le vecteur exprimant la distance Euclidienne entre le centre de gravité d'un objet et un point *SURF* appartenant à un bloc β_i. Chaque bloc β_i comprenant un nombre n_i de points *SURF*, avec des vecteurs $(\tau_{i1}, \tau_{i2}, \ldots, \tau_{in_i},)$, la moyenne de la distance Euclidienne entre l'ensemble des points *SURF* par bloc peut être notée par γ_i telle que :

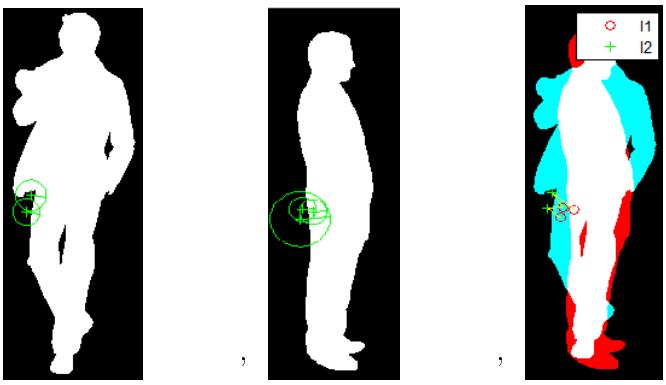
$$\gamma_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \tau_{ij}. \tag{3.35}$$

On considère la valeur Λ_i comme la distance entre les valeurs γ_i^r et γ_i^t de deux blocs homologues telle que $\Lambda_i = (\gamma_i^r - \gamma_i^t)^2$. Ces deux blocs homologues appartiennent

à un objet test (*request*, r) et un objet dans la base de donnée (*target*, t). La mise en correspondance par blocs est effectuée par une opération de pondération des scores ψ_i par les distances Euclidiennes Λ_i. Elle s'effectue par le biais d'une comparaison des scores qui sont calculés pour deux objets quelconques suivant l'équation 3.36.

$$\mathcal{SCORE}_{final} = \sum_{i=1}^{|\mathcal{B}|} \psi_i \times \Lambda_i \tag{3.36}$$

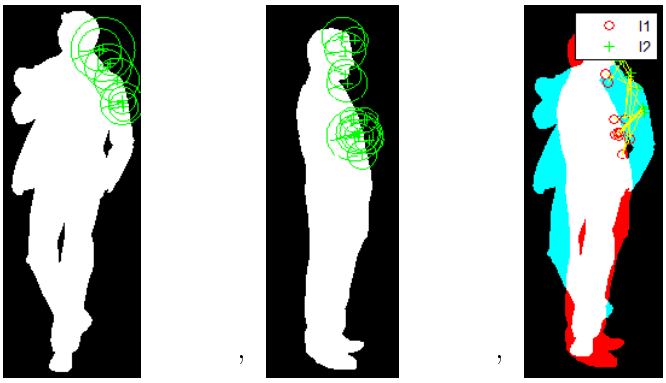
Les figures 3.16 et 3.17 suivantes illustrent des points d'intérêt sélectionnés dans des blocs différents ainsi que leur mise en correspondance. Les deux premières images illustrent deux objets tests A et B avec des points *SURF* localisés au niveau du bloc 4 et du bloc 7, conformément à la subdivision de la figure 3.15. La troisième image illustre la mise en correspondance des points d'intérêt détectés dans les blocs homologues.



(a) Image A (b) Image B (c) A vs B

FIGURE 3.16 – Correspondance des points SURF - bloc 4.

Les images a) et b) sont des images tests. L'image c) présente l'appariement des points d'intérêt du bloc 4 des images tests.



(a) Image A (b) Image B (c) A vs B

FIGURE 3.17 – Correspondance des points SURF - bloc 7.

Les images a) et b) sont des images tests. L'image c) présente l'appariement des points d'intérêt du bloc 7 des images tests.

3.6 Représentation de l'ensemble d'apprentissage

La représentation de l'ensemble d'apprentissage pour nos algorithmes a été effectuée à travers la création d'une ontologie des objets ainsi que de leurs poses et actions. Une ontologie définit un ensemble de concepts, leurs caractéristiques, ainsi que les relations qui peuvent exister entre les concepts [17]. Elle permet de décrire de façon formelle la sémantique des objets par la création d'une base de connaissance. La création d'une ontologie est un processus itératif, dans lequel la base de connaissance est enrichie au fil du temps. De façon formelle, considérons l'ensemble de concepts $\mathcal{C}$ présents dans notre ontologie. Les éléments $\mathcal{C}_i$ de cet ensemble peuvent être liés par des relations d'ordre partielles $\preceq$, telles que $\forall(C_1, C_2) \in \mathcal{C}^2, C_1 \preceq C_2$ indique que C_1 est un sous-concept de C_2. La relation d'ordre partielle exprime les liens orientés entre les concepts qui

sont représentés dans notre ontologie. Nous pouvons également définir un ensemble $\mathcal{T}$ d'attributs qui peuvent être associés aux différents concepts de notre ontologie.

La représentation de l'ensemble d'apprentissage permet de créer une hiérarchie des objets, de leurs poses et actions, pour constituer une base de connaissance. Dans le niveau hiérarchique le plus élevé de notre ontologie, nous y retrouvons le concept d'*objet mobiles*. Ce niveau hiérarchique permet de distinguer les sous-concept d'*objets simples* et d'*objets composites* (ex. silhouettes de chien accompagné de son maître). Ces *objets composites* ne seront pas traités dans ce travail. Le sous-concept d'*objets simples* regroupe les classes d'objets (*Chiens, Humains, Voitures, etc.*). Les attributs de notre ensemble $\mathcal{T}$ telles que les *poses gauche ou droite* et les *actions* seront associés aux concepts définissant les objets manipulés par les algorithmes. Un objet de la classe *Chiens* peut par exemple être dans une pose gauche ou dans une pose droite. La figure 3.18 donne une illustration de la représentation conceptuelle de notre base de connaissance.

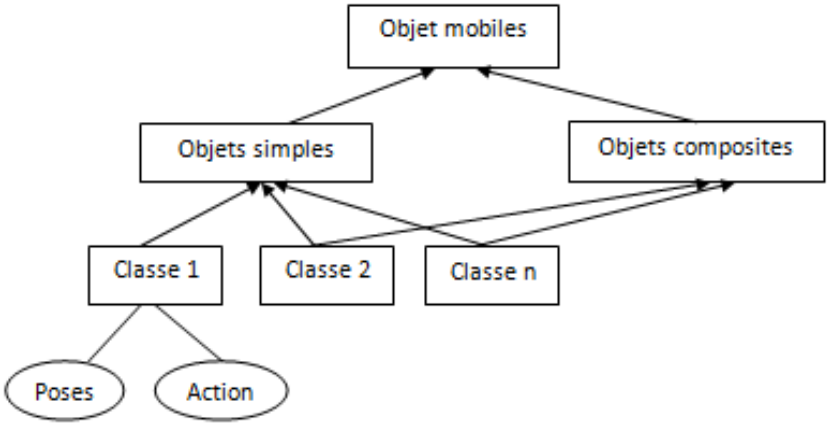


FIGURE 3.18 – Ontologie des objets.

Chapitre 4

Expérimentation

Dans ce chapitre, nous allons présenter les différents tests que nous avons effectués concernant nos deux méthodes proposées pour la reconnaissance des objets en mouvement dans les vidéos ; ces deux méthodes étant la reconnaissance par descripteur global (forme contextuelle) et la reconnaissance par les points d'intérêt (mise en correspondance par les points d'intérêt). Nous présenterons également les résultats que nous avons obtenus. Notre expérimentation se déroulera selon les points énumérés ci-dessous :

1. Élaboration de la base de données annotée
2. Établir des critères d'évaluation des résultats de reconnaissance
3. Reconnaissance par descripteur global et résultats
4. Reconnaissance par descripteur local et résultats
5. Conclusion

4.1 Élaboration de la base de données

Dans le cadre de notre série de tests, nous avons créé une base de données d'objets de trois classes : *Chiens, Humains, Voitures*. Chaque classe d'objets contient 70 individus. Les objets composants cette base de données représentent les objets connus à partir desquels nous identifierons les objets en mouvement qui sont extraits des vidéos.

Notre base de données d'objets est structurée selon l'ontologie présentée dans le chapitre précédent. Il s'agit dans ce cas d'une représentation partielle des objets ainsi que leur pose, selon notre ontologie de la section 3.6 du chapitre précédent. Dans ce sens, cette base de données est composée de trois classes d'objets (*Chiens, Humains et Voitures*), et les objets de chaque sont associés à une pose. La pose permet de distinguer les objets regardant vers la droite de ceux regardant vers la gauche. La structure de notre base de données d'objets est illustrée sur la figure 4.1.

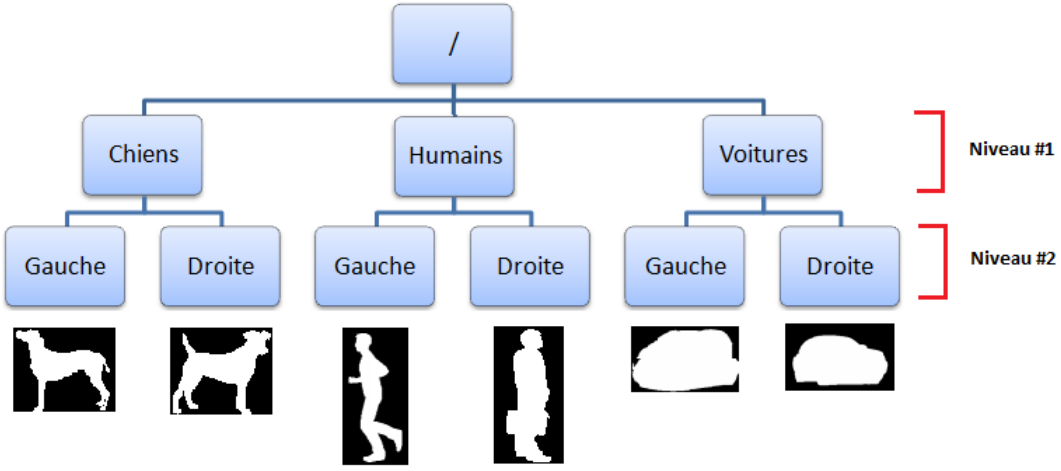


FIGURE 4.1 – Base de données d'objets.

4.2 Les jeux de tests

Nos tests sont constitués d'instances d'objets appartenant aux différentes classes. Chaque jeu de test comprend 20 objets, choisis aléatoirement. Nous précisons que les objets constituant à la fois notre base de données et nos jeux de tests sont issus des images binaires obtenus à la fin du processus de pré-traitement (c-à-d processus de sous-

traction de fond). Ces objets représentent des silhouettes d'individus dans leur classes, par exemple un seul chien, un seul humain.

La figure 4.2 montre quelques individus de notre jeu de tests. Les trois classes d'objets (*Chiens, Humains, Voitures*) y sont présentes. Les deux premières colonnes de chaque ligne montrent des objets dans une pose droite et les deux dernières colonnes présentent des objets dans une pose gauche.

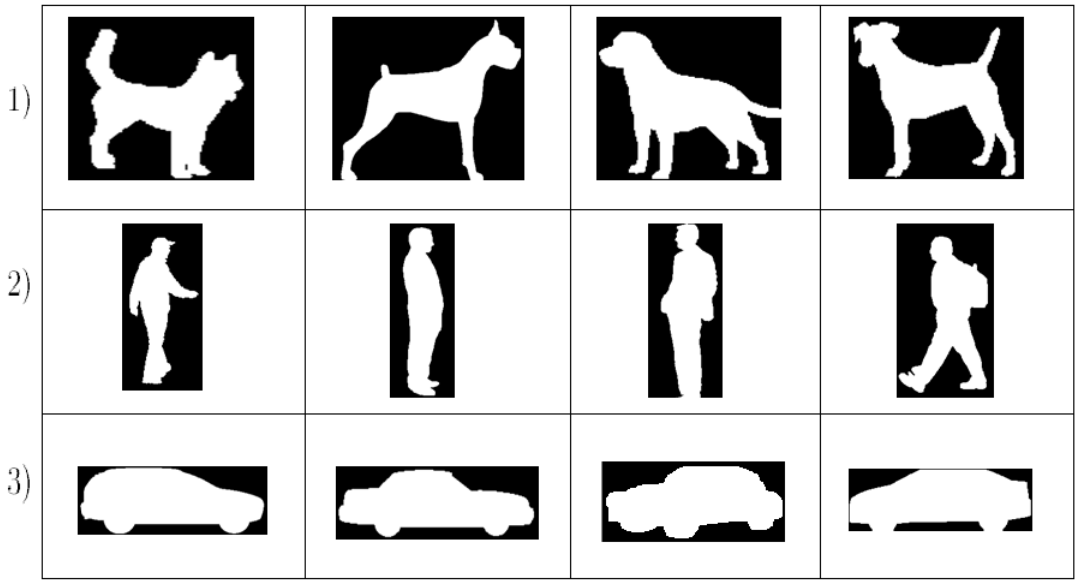


FIGURE 4.2 – Images tests dans une pose droite ou gauche. 1)-Chiens, 2)-Humains, 3)-Voitures

.

4.3 Critères d'évaluation des tests

Les tests de reconnaissance d'objets se dérouleront sur deux niveaux (voir figure 4.1) :

- le premier niveau est relatif à la reconnaissance de la classe de l'objet et permettra de déterminer si l'objet est un chien, une voiture ou un humain.
- le second niveau est relatif à la reconnaissance de la pose de l'objet et permettra de déterminer si l'objet est dans une pose gauche ou dans une pose droite.

L'évaluation des méthodes de reconnaissance est une étape importante dans le sens où elle permet de mesurer les performances du système. Des méthodes existantes permettent d'évaluer la performance de nos méthodes de reconnaissance. Ainsi, pour la reconnaissance des objets, nous évaluerons notre système par le nombre d'objets de tests qui seront correctement identifiés par rapport au nombre de requêtes envoyées au système. Pour les systèmes de reconnaissance basés sur la méthode des *KPPV*, la bonne identification d'une requête a une relation directe avec la *précision* de la recherche dans la base de données. Par définition, la *précision* indique le nombre d'objets pertinents par rapport au nombre d'objets retournés par le système. Ainsi, lorsque la *précision* de la recherche est supérieure à 50%, cela signifie que le nombre d'objets dans la base de données, de même classe que celui de la requête (et qui sont considérés par le *KPPV*) est suffisant pour identifier correctement l'objet.

4.3.1 Matrice de confusion

Pour évaluer la reconnaissance des objets pour chaque type de descripteur avec l'algorithme des *KPPV*, nous présenterons les matrices de confusion pour les trois classes d'objets (*Chiens, Humains, Voitures*). La matrice de confusion permet de rapporter le pourcentage d'objets correctement identifiés pour chaque classe. Elle renseigne également sur les confusions du système entre certaines classes d'objets.

4.3.2 Influence de la valeur de k pour les KPPV

La reconnaissance des objets avec la méthode des KPPV dépend des valeurs qui sont attribuées au paramètre k. Cette valeur dépend du nombre d'objets de chaque classe qui sont présents dans la base de données. D'une manière générale, pour identifier un objet d'une classe, la valeur attribuée au paramètre k doit être inférieure au nombre total d'objets de cette classe dans la base de données. Si on considère que n représente le nombre total d'objets d'une classe dans la base de données, la valeur maximale de k est définie par la formule $k \leq 2n - 1$. Pour évaluer les résultats d'expérimentation par la méthode des *K Plus Proches Voisins*, nous avons varié la valeur de k de 1 à 11, puis nous avons relevé pour chacune des valeurs de k, l'évolution du nombre d'objets tests correctement identifiés par le système. La figure 4.3 montre la courbe d'évolution du nombre d'objets tests correctement identifiés par rapport aux différentes valeurs de k, pour la reconnaissance de classe.

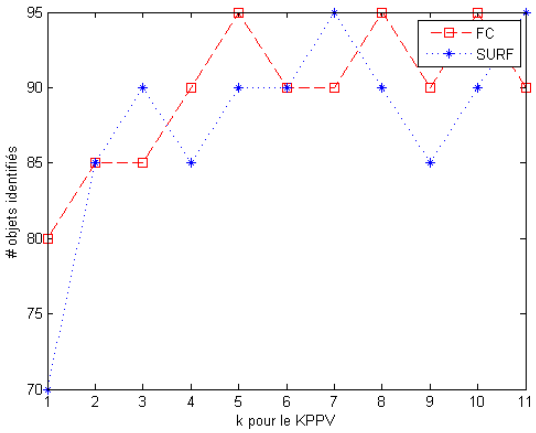


FIGURE 4.3 – Courbe d'évolution de k et du nombre d'objets identifiés.

On remarque sur cette figure que le nombre d'objets tests correctement identifiés a une stabilité autour de 95% pour $k \geq 5$.

4.4 Résultats d'expérimentation

Nous avons procédé à l'expérimentation de nos méthodes de reconnaissance pour les trois classes d'objets en utilisant nos exemples de tests mentionnées ci-dessus ainsi que la méthode d'évaluation choisie. Dans un premier temps nous présentons les résultats d'expérimentation avec la forme contextuelle (FC). Nous présentons ensuite les résultats d'expérimentation avec les points d'intérêt ainsi que par la méthode des *Bag of Features*.

4.4.1 Reconnaissance d'objets avec le descripteur global

4.4.1.1 Application de la méthode des KPPV

La matrice de confusion pour la reconnaissance de la classe des objets par la méthode des *KPPV* est illustrée par le tableau ci-dessous.

	Chiens	**Humains**	**Voitures**
Chiens	0,90	0,05	0,05
Humains	0,05	0,95	0
Voitures	0,10	0	0,90

TABLE 4.1 – Reconnaissance de classe - descripteur global.

Les valeurs rapportées dans la diagonale du tableau 4.1 étant relativement élevées, cela traduit la qualité de la représentation des objets par notre descripteur global (la

forme contextuelle) surtout dans la reconnaissance des classes d'objets. Les différences visuelles qui peuvent être observées à travers la forme géométrique des objets n'influent pas de façon significative sur leur description par la forme contextuelle. Ces valeurs indiquent également que la forme contextuelle est plus sensible aux objets des classes *Chiens* et *Voitures*. Cela peut se justifier par la forme géométrique des objets de ces deux classes dont la longueur est plus importante que la hauteur.

Les matrices de confusion pour la reconnaissance avec la forme contextuelle des poses gauche (g) et droite (d) des objets sont illustrées par les tableaux ci-dessous.

	Chiens		**Humains**		**Voitures**	
Chiens	0,70 (g)	0,10 (d)	0,05		0,15	
Humains	0,05		0,75 (g)	0,10 (d)	0,10	
Voitures	0,15		0,05		0,75 (g)	0,05 (d)

TABLE 4.2 – Reconnaissance de la pose gauche (g) - descripteur global.

	Chiens		**Humains**		**Voitures**	
Chiens	0,05 (g)	0,70 (d)	0,1		0,15	
Humains	0,15		0,05 (g)	0,70 (d)	0,10	
Voitures	0,20		0		0,05 (g)	0,75 (d)

TABLE 4.3 – Reconnaissance de la pose droite (d) - descripteur global.

L'analyse des pourcentages présentés dans les diagonales de ces tableaux révèle une diminution du nombre d'objets correctement identifiés par le système. Cette baisse traduit une plus grande sensibilité de notre descripteur global par rapport aux changements

dans la position des objets. La baisse est plus remarquée au niveau de la pose droite des chiens et des humains. Cela est essentiellement dû à la forme géométrique des objets dans cette sous-catégorie dans la base de données. Ces objets, en général présentent des déformations beaucoup plus marquées comparativement aux objets qui sont dans une pose gauche.

4.4.1.2 Application de la méthode naïve de Bayes

Des tests ont été effectués pour la forme contextuelle avec la méthode naïve de Bayes. Les résultats de reconnaissance par l'application de cette méthode sont illustrés dans les tableaux 4.4 et 4.5 sous forme de pourcentages. Le tableau 4.4 présente les résultats qui sont relatifs à la reconnaissance des classes d'objets (premier niveau) et le tableau 4.5 présente les résultats de reconnaissance pour le second niveau.

Chiens	Humains	Voitures
0,95	0,96	0,89

TABLE 4.4 – Méthode naïve de Bayes - FC - Reconnaissance de classe.

	Pose gauche	Pose droite
Chiens	0,79	0,76
Humains	0,80	0,82
Voitures	0,83	0,80

TABLE 4.5 – Méthode naïve de Bayes - FC - Reconnaissance des poses.

On remarquera d'après les pourcentages de reconnaissance rapportés dans le tableau 4.4 que la méthode naïve de Bayes identifie les objets des classes *Chiens* et *Humains* avec plus de 90% de réussite. Le pourcentage rapporté pour la classe *Voitures*

est légèrement en dessous de 90%. Cela confirme un des avantages de cette méthode statistique qui, malgré la forte hypothèse d'indépendance des attributs des descripteurs d'objets, offre des pourcentages de reconnaissance élevés et est souvent utilisée en comparaison avec les autres méthodes statistiques existantes. En revanche, on observe une baisse très importante, pour les objets de la classe *Chiens*, relativement à la reconnaissance de la pose des objets. Nous pouvons cependant déduire d'après ce résultat que la reconnaissance de la pose avec la méthode naïve de Bayes est meilleure par rapport à celle de la méthode des $KPPV$ pour la classe des humains.

4.4.2 Reconnaissance d'objets avec les descripteurs locaux

Pour effectuer les tests avec les points d'intérêt, nous avons fixé un certain seuil relatif au nombre de descripteurs $SURF$ nécessaires pour chaque objet. Ainsi, chaque objet doit avoir au minimum 10 points d'intérêt pour être traité par l'algorithme de reconnaissance. Cela nous permet de renforcer la crédibilité des résultats de reconnaissance. Le nombre de points d'intérêt détectés pour chaque objet varie avec la résolution de l'image. Les images haute résolution génèrent plus de points d'intérêt que les images basse résolution. Nous avons augmenté la résolution de certaines de nos images tests afin de détecter plus de points d'intérêt et d'avoir le nombre minimum de descripteurs $SURF$ requis.

4.4.2.1 Application de la mise en correspondance

	Chiens	**Humains**	**Voitures**
Chiens	0,95	0	0,05
Humains	0	0,95	0,05
Voitures	0	0,1	0,9

TABLE 4.6 – Reconnaissance de classe - SURF.

D'après le tableau 4.6 , la reconnaissance de la classe des objets par la méthode de mise en correspondance montre des pourcentages d'objets correctement identifiés plus élevés pour les classes *Humains* et *Chiens* par rapport à la classe *Voitures*. Ces valeurs sont comparables à celles observées pour la reconnaissance des classes avec le descripteur global (FC). Cependant, la valeur reportée dans ce tableau est légèrement inférieure pour la classe *Voitures*, en comparaison avec celle reportée dans le tableau 4.1 pour la forme contextuelle.

	Chiens		**Humains**		**Voitures**	
Chiens	0,75 (g)	0,10 (d)	0,05		0,1	
Humains	0,10		0,75 (g)	0,10 (d)	0,05	
Voitures	0,25		0		0,7 (g)	0,05 (d)

TABLE 4.7 – Reconnaissance de la pose gauche (g) - SURF.

	Chiens		**Humains**		**Voitures**	
Chiens	0,10 (g)	0,75 (d)	0		0,15	
Humains	0,15		0,05 (g)	0,80 (d)	0	
Voitures	0,15		0,05		0,05 (g)	0,75 (d)

TABLE 4.8 – Reconnaissance de la pose droite (d) - SURF.

Les valeurs reportées pour la reconnaissance des poses avec notre méthode de mise en correspondance des points d'intérêt montrent que dans la majorité des cas, certains objets de la classe *Voitures* sont identifiés comme appartenant à la classe *Chiens*. Cette tendance révèle l'importance de la forme géométrique des objets appartenant à ces deux classes. En effet les objets de ces deux classes ont une forme géométrique presque similaire, leur largeur étant plus grande que leur hauteur. La reconnaissance de la pose des objets appartenant à la classe des humains par contre, ne présente pas cette tendance. Les valeurs des tableaux 4.7 et 4.8 traduisent le fait que la mise en correspondance est également sensible à la pose des objets des différentes classes.

La reconnaissance des classes d'objets avec cette méthode montre que les objets de la classe *Humains* ont été mieux identifiés comparativement aux objets des deux autres classes. Ce résultat indique que la division par blocs des objets favorise en particulier les objets de la classe *Humains*. Les valeurs rapportées pour la reconnaissance des objets de la classe *Voitures* sont inférieures à celles des objets des autres classes et traduisent la sensibilité de notre méthode à la forme géométrique de la silhouette des objets de cette classe qui est peu déformable.

4.4.3 Problèmes posés

Les différents résultats des expérimentations qui sont rapportés dans les tableaux des sections précédentes montrent des pourcentages élevés pour la reconnaissance des classes d'objets. Cependant, la silhouette de certains objets utilisés lors des tests ne permet pas d'effectuer une nette distinction entre les classes. Nous illustrons cette remarque par les objets de la figure 4.4 ci-dessous. On remarquera sur cette figure que les silhouettes des chiens ont l'apparence d'avoir une position droite qui peut être semblable aux silhouettes d'humains. De la même manière les silhouettes des humains ont une envergure qui peut géométriquement être comparée à l'envergure des silhouettes d'objets de la classe *Chiens* ou de la classe *Voitures*.

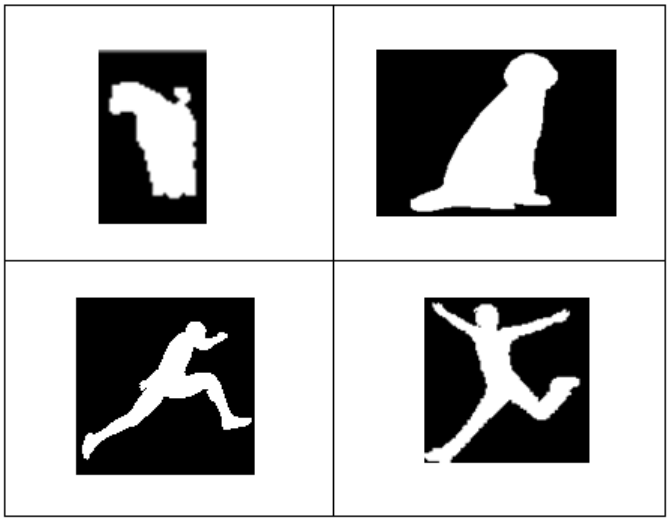


FIGURE 4.4 – Images tests créant une ambiguïté. En haut, chiens identifiés comme des humains. En bas, humains identifiés comme des voitures ou des chiens.

4.4.4 Représentation d'objets par les deux descripteurs

Les expérimentations ont révélé que les formes géométriques d'objets appartenant aux trois classes (*Chiens, Humains et Voitures)* ont été représentées avec fiabilité par les deux types de descripteurs (descripteur global et descripteurs locaux). Ces objets ont été correctement identifiés par le système. La première rangée de la figure 4.5 ci-dessous, montre des exemples d'objets des trois classes correctement représentés par les deux descripteurs.

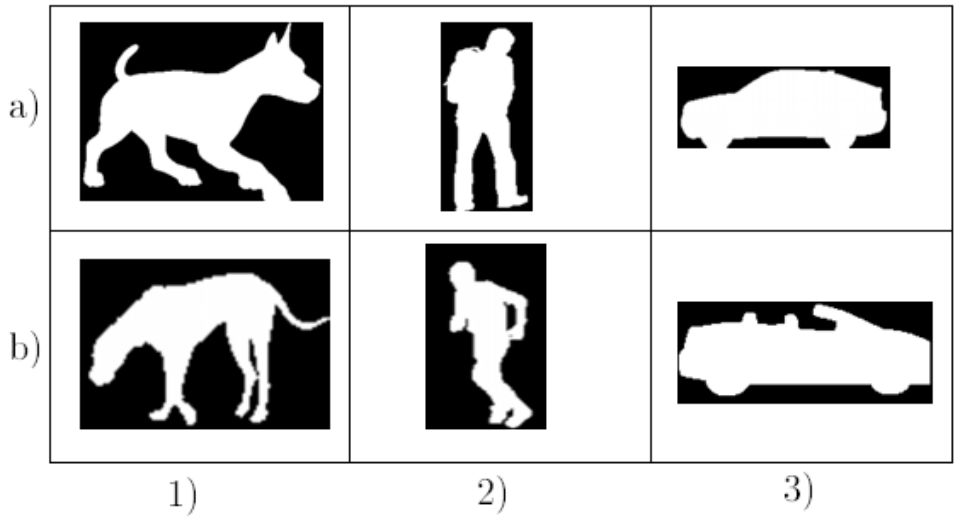


FIGURE 4.5 – Objets représentés avec fiabilité par les deux descripteurs. 1) Chiens, 2) Humains, 3) Voitures.

Les tests ont également révélé des disparités dans la description des objets. Ces disparités ont entraîné une confusion de classe de la part du système. La deuxième rangée de la figure 4.5 montre un exemple d'objets de chaque classe dont la représentation par les deux types de descripteurs a conduit à une erreur du système.

4.4.5 Méthode comparative (sacs de mots visuels)

Nous avons également effectué la reconnaissance des classes d'objets (premier niveau) par une méthode comparative : celle des *Bag oF Features (BoF)*. La méthode de reconnaissance par la méthode des *BoF* s'effectue en plusieurs étapes dont l'une a pour objectif de construire un dictionnaire de "mots visuels" des objets dans le but de les identifier. Les images binaires issues du processus de pré-traitement et contenant des objets identifiables, peuvent être modélisées par des *sacs de mots visuels ou (Bag of Features)* [26] regroupant leurs points d'intérêt. Le processus de reconnaissance se déroule selon les étapes suivantes :

1. *Détection et description des points d'intérêt :*

   L'étape de détection et de description des points d'intérêt s'effectue par les détecteurs existant, en l'occurrence le *SURF*.

2. *Création du dictionnaire de mots visuels :*

   La création du dictionnaire de mots visuels par utilisation de la méthode de partitionnement des *k-moyennes* pour créer des groupes de descripteurs *SURF* en fonction d'une métrique de distance. Chaque descripteur *SURF* est alors affecté à une partition en fonction de la métrique de distance utilisée. La méthode des *k-moyennes* [16] permet de regrouper les descripteurs d'objets en différentes partitions. Le processus d'affectation des descripteurs aux différentes partitions est un processus récursif dans lequel les moyennes sont mises à jour après ajout de chaque nouveau descripteur. Les centres des partitions $\mathcal{K}_c$ constituent la moyenne des descripteurs. Le dictionnaire de mots visuels créé durant cette étape est constitué par l'ensemble des $\mathcal{K}_c$.

3. *Représentation des objets :*

   Les objets à identifier sont ensuite représentés par des histogrammes dont les entrées représentent les centres $\mathcal{K}_c$. Ces histogrammes permettent de quantifier les descripteurs par objet.

4. *Apprentissage et reconnaissance :*

   La dernière étape consiste à utiliser un algorithme de reconnaissance pour affecter chaque objet identifiable à sa classe. Pour ce faire, nous construirons des modèles d'apprentissage et nous appliquerons la méthode des *KPPV* pour identifier les objets.

La description des objets par les sacs de mots visuels ne prend pas en compte les caractéristiques géométriques de ces objets. Cette expérimentation avec la méthode des *Bag of Features* a pour objectif de comparer les résultats obtenus avec ceux des tests avec notre méthode de mise en correspondance. La matrice de confusion pour la reconnaissance des classes d'objets avec les *BoF*, est illustrée par le tableau 4.9.

	Chiens	**Humains**	**Voitures**
Chiens	0,95	0,05	0
Humains	0,05	0,8	0,15
Voitures	0,05	0,1	0,85

TABLE 4.9 – Reconnaissance de classe - BoF.

On remarque d'après les valeurs reportées que les objets des classes *Chiens* et *Voitures* ont été relativement mieux identifiés par rapport aux objets de la classe *Humains*. Cela traduit le fait que les *BoF* qui sont créés pour les silhouettes de chiens et de voitures

favorisent une meilleure discrimination des objets de ces deux classes. En comparaison avec la mise en correspondance, les pourcentages d'objets correctement identifiés, rapportés pour les classes *Humains* et *Voitures* avec cette méthode de description, sont inférieurs à ceux obtenus avec la méthode de mise en correspondance (voir tableau 4.6). Cette tendance confirme l'influence positive de la prise en compte de la géométrie des objets dans le processus de reconnaissance avec les points d'intérêt.

4.5 Sommaire du chapitre

Les tests qui ont été effectués avec les deux descripteurs (forme contextuelle et points d'intérêt) ont donné différents résultats pour les deux niveaux de reconnaissance (reconnaissance de classe et reconnaissance de pose). Les tests de reconnaissance des classes d'objets (premier niveau) par leur forme contextuelle ont donné de meilleurs résultats pour les classes *Chiens* et *Humains*, par rapport à la classe *Voitures*, même si les pourcentages d'objets correctement identifiés qui sont obtenus pour la classe *Voitures* sont proches de 90%. La reconnaissance du premier niveau avec les descripteurs locaux (points d'intérêt), montre également des pourcentages élevés pour les trois classes d'objets. En considérant la forme géométrique variable des objets, les résultats de la reconnaissance de classes permettent d'affirmer que les méthodes proposées répondent aux propriétés d'invariance telles que l'échelle et la translation.

Les tests effectués pour le second niveau (*Reconnaissance de poses d'objets*) ont donné des pourcentages d'objets correctement identifiés inférieurs à 80%, aussi bien pour la forme contextuelle que pour les points d'intérêt. Ces résultats traduisent une grande sensibilité des deux types de descripteurs utilisés (descripteur global et descripteurs lo-

caux) par rapport aux différents changements de pose des objets. Ainsi, ces changements de pose d'objets, entraînent une plus grande déformation des objets et influent de façon négative sur leur représentation par les descripteurs.

Chapitre 5

Conclusion générale

Dans ce travail, nous avons présenté différentes approches de reconnaissance d'objets en mouvement dans les vidéos. Dans une première étape des techniques d'extraction d'objets du mouvement ont été utilisées afin de détecter et d'isoler les objets en mouvement dans les vidéos. Dans une seconde etape, nous avons proposé deux approches de reconnaissance d'objets qui sont basées sur deux types de descripteurs : la forme contextuelle et les points d'intérêt *SURF*.

Pour valider nos solutions, nous avons effectué plusieurs expérimentations avec deux méthodes statistiques qui sont le *K Plus Proches Voisins* et la méthode naïve de Bayes. La méthode naïve de Bayes a seulement été appliquée pour la reconnaissance avec la forme contextuelle. Cette contrainte se justifie par la dimension assez élevée des descripteurs *SURF* par objet. En effet, nous avons un vecteur de dimension $d = 64$ par point d'intérêt *SURF* et chaque objet de notre système génère en moyenne $d = 30$ points d'intérêt *SURF*. Il s'en suit alors, en moyenne pour chaque objet des descripteurs de dimension $d = 1920$, ce qui a une influence négative sur les résultats de la reconnaissance avec la méthode naïve de Bayes.

Les principales contributions de notre travail sont les suivantes :

1. **La reconnaissance multi-niveaux et multi-classe d'objets**

Notre travail permet d'identifier les objets en mouvement suivant deux niveaux. Le premier niveau de reconnaissance concerne la classe de l'objet (*Chiens, Humains, Voitures*) et le second niveau concerne la pose de l'objet (pose gauche ou pose droite).

2. **Nouvelle version du *shape context* dénommée *forme contextuelle***

   Nous avons crée une nouvelle version du *shape context* orignal de Belongie *et al.* [4] dans l'objectif de réduire la dimension du descripteur et d'améliorer le calcul des distances entre les points du contour et le centre de gravité des objets. Nous avons remplacé la distance euclidienne par la distance de *Mahalanobis* qui est plus adaptée à la forme géométrique des objets.

3. **Nouvelle représentation des objets avec les descripteurs *SURF***

   La mise en correspondance des points d'intérêt, sans tenir compte des caractéristiques géométriques des objets, entraîne des pourcentages d'objets correctement identifiés très faibles. Cela est en grande partie dû à la corrélation parfaite qui peut être obtenue entre les descripteurs *SURF* d'objets appartenant à différentes classes, et schématisée sur la figure 3.14 du chapitre Méthodologie. En effet, les points d'intérêt *SURF* qui sont détectés sur une image binaire contenant un objet en mouvement sont caractéristiques des zones de variation d'intensité. Ces zones sont liées aux caractéristiques géométriques locales des objets. Nous avons donc introduit une nouvelle représentation par bloc des objets permettant de mesurer leur degré de similarité par groupe de points d'intérêt. Cette représentation permet d'effectuer la mise en correspondance par groupes de points d'intérêt. Nous avons également pris en compte la distribution géométrique relative des points d'intérêt *SURF* par rapport au centre de gravité des objets.

4. Création d'une base de données multi-classe et multi-niveaux annotée

   Notre base de données est composée d'objets de trois classes : *Chiens, Humains et Voitures*. Les individus de chaque classe sont annotés selon leur pose gauche ou droite.

5. Les résultats d'expérimentation et la performance de notre méthode montrent des pourcentages de reconnaissance assez élevés (au-delà de 90% en moyenne) pour le premier niveau de reconnaissance. Ces résultats encourageants montrent que notre solution de reconnaissance d'objets peut être déployée dans un scénario de vidéo surveillance.

Nous relevons certaines améliorations qui peuvent être apportées à notre travail ainsi que certaines perspectives. Ils portent sur les points suivants :

1. L'amélioration de l'extraction d'objets en mouvement dans les vidéos avec la possibilité d'isoler non seulement des silhouettes uniques (simples) mais également celle de pouvoir diviser les silhouettes d'objets composites en individus séparés. Il s'agira par exemple de diviser une silhouette représentant un chien lié à un humain en deux silhouettes différentes.
2. L'amélioration des techniques de détection de changement pour arriver à isoler des objets identifiables et aussi éliminer les ombres qui les accompagnent.
3. L'extension de nos approches de description d'objets à d'autres classes. Il s'agit d'une piste possible à explorer et qui pourra aboutir à la description fiable d'objets appartenant à d'autres classes ou à la création d'un nouveau type de descripteur d'objets.

Bibliographie

[1] M.S. Allili and D. Ziou. Object of interest segmentation and tracking by using feature selection and active contours. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[2] M.S. Allili, D. Ziou, N. Bouguila and S. Boutemedjet. Image and video segmentation by combining unsupervised generalized gaussian mixture modeling and feature selection. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(10) :1373–1377, 2010.

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

[4] S. Belongie, J. Malik and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4) :509–522, 2002.

[5] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, pages 424–430. Springer-Verlag New York, First Edition, 2006.

[6] A. Bovik. *Handbook of Image and Video Processing*, pages 393–438. Elsevier Academic Press, 2nd Edition, 2005.

[7] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. *IEEE Int'l Conference on Computer Vision*, 1 :105–112, 2001.

[8] G. J. Brostow, J. Fauqueur and R. Cipolla. Semantic object classes in video : A high-definition ground truth database. *Pattern Recognition Letters*, 20(2) :88–97, 2009.

[9] G. J. Brostow, J. Shotton, J. Fauqueur and R. Cipolla. Segmentation and recognition using structure from motion point clouds. *European Conference on Computer Vision*, pages 44–57, 2008.

[10] T. Brox, L. Bourdev, S. Maji and J. Malik. Object segmentation by alignment of poselet activations to image contours. *IEEE Conference on Computer vision and Pattern Recognition*, pages 2225–2232, 2011.

[11] S-W. Chein, L.K. Wang and J-H. Lan. Real-time background subtraction for video surveillance : From research to reality. *IEEE Int'l Colloquium on Signal Processing and Its Applications*, pages 1–6, 2011.

[12] D. Comaniciu, P. Meer and Senior Member. Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5) :603–619, 2002.

[13] D. Cremers, S. Osher and S. Soatto. Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *Int'l Journal of Computer Vision*, 69(3) :335–351, 2006.

[14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1 :886–893, 2005.

[15] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros and M. Hebert. An empirical study of context in object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278, 2009.

[16] R.O. Duda, P. E. Hart and D. G. Stork. *Pattern Classification.* John Wiley & Sons, 2000.

[17] N. Durand, S. Derivaux, G. Forestier, C. Wemmert, P. Gancarski, O. Boussaid and A. Puissant. Ontology-based object recognition for remote sensing image interpretation. *IEEE 19th Int'l Conference on Tools with Artificial Intelligence*, 1 :472–479, 2007.

[18] A. Elgammal, D. Harwood and L. Davis. Non-parametric model for background subtraction. *IEEE Frame-Rate Workshop*, pages 751–767, 2000.

[19] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan. Visual object detection with deformable part models. *Communications of the ACM*, 56(9) :97–105, 2013.

[20] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int'l Journal of Computer Vision*, 61(1) :55–79, 2005.

[21] R.C. Gonzalvez and R. E. Woods. *Digital Image Processing.* Prentice Hall, 3rd Edition, 2008.

[22] C. Gu, J. J. Lim, P. Arbeláez and J. Malik. Recognition using regions. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037, 2009.

[23] R. Haccoun and D. Cousineau. *Statistiques : concepts et applications*, pages 149–152. Les presses de l'Université de Montréal, 2007.

[24] H. Harzallah, F. Jurie and C. Schmid. Combining efficient object localization and image classification. *IEEE Int'l Conference on Computer Vision*, pages 237–244, 2009.

[25] H. Jia and Y. Zhang. Fast human detection by boosting histograms of oriented gradients. *IEEE Int'l Conference on Image and Graphics*, pages 683–688, 2007.

[26] Y. Jiang, C. Ngo and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *Proceedings of the 6th ACM Int'l Conference on Image and Video Retrieval*, pages 494–501, 2007.

[27] T. Joachims. Text categorization with support vector machines : Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, 1998.

[28] Z. K. Ku, C. F. Ng and S. W. Khor. Shape-based recognition and classification for common objects - an application in video scene analysis. *IEEE Int'l Conference on Computer Engineering and Technology*, 3 :13–16, 2010.

[29] G. Larivière. Segmentation d'objets : Approches par analyse de forme et apprentissage probabiliste. Master's thesis, Université du Québec en Outaouais, 2012.

[30] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image and Vision Computing*, 27(5) :523–534, 2009.

[31] S. H. Lee, S. Sharma, L. Sang, J. Park and Y. G. Park. An intelligent video security system using object tracking and shape recognition. *Int'l Conference on Advanced Concepts for Intelligent Vision Systems*, pages 471–482, 2011.

[32] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. *British Machine Vision Conference*, pages 759–768, 2003.

[33] G. Li, Y. Wang and W. Shu. Real-time moving object detection for video monitoring systems. *IEEE Int'l Symposium on Intelligent Information Technology Application*, 3 :163–166, 2008.

[34] K-L. Lim and H. Kiani Galoogahi. Shape classification using local and global features. *IEEE Pacific-Rim Symposium on Image and Video Technology*, pages 115–120, 2010.

[35] A.J. Lipton, H. Fujiyoshi and R.S. Patil. Moving target classification and tracking from real-time video. *IEEE Workshop on Applications of Computer Vision*, pages 8–14, 1998.

[36] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int'l Journal of Computer Vision*, 1(60) :63–86, 2004.

[37] J. C. Nascimento and J. S. Marques. Performance evaluation for object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4) :761–774, 2006.

[38] M. Oren, C. Papageorgiou and P. Sinha. Pedestrian detection using wavelet templates. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–199, 1997.

[39] N. Paragios. A level set approach for shape-driven segmentation and tracking of the left ventricle. *IEEE Transactions on Medical Imaging*, 22(6) :773–776, 2003.

[40] P.J. Pedrocca and M.S. Allili. Real-time people detection in videos using geometrical features and adaptive boosting. *Int'l Conference on Image Analysis and Recognition*, LNCS 6753 :314–324, 2011.

[41] M. Piccardi. Background subtraction techniques : A review. *IEEE Int'l Conference on Systems, Man and Cybernetics*, 4 :3099–3104, 2004.

[42] T. Poggio. Marr's computational approach to vision. *Trends in Neurosciences*, 4 :258–262, 1981.

[43] J. Ricard, D. Coeurjolly and A. Baskurt. Generalization of angular radial transform. *IEEE Int'l Conference on Image Processing*, 4 :2211–2214, 2004.

[44] M. Rousson and N. Paragios. Shape priors for level set representations. *European Conference on Computer Vision*, LNCS 2351 :78–92, 2002.

[45] J. Russell and R. Cohn. *Pandemonium Architecture*. Google On Demand Book, 2012.

[46] C. Schmid, R. Mohr and C. Bauckhage. Evaluation of interest point detectors. *Int'l Journal of Computer Vision*, 37(2) :151–172, 2000.

[47] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. *IEEE Int'l Workshop on Content-based Access of Image and Video Databases*, pages 61–71, 1998.

[48] S. M. Smith and J. M. Brady. Susan - A new approach to low level image processing. *Int'l Journal of Computer Vision*, 23(34) :45–78, 1997.

[49] L. Squire, F.E. Bloom, N.C. Spitzer, L.R. Squire, D.Berg, S. du Lac and A. Ghosh. *Fundamental Neuroscience*. Academic Press, 2008.

[50] C. Stauffer and W.E.L Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8) :747–757, 2000.

[51] E. A. Styles. *Attention, Perception And Memory - An Integrated Introduction*. Psychology Press, 2006.

[52] Y. Su, R. Qian and Z. Ji. Surveillance video sequence segmentation based on moving object detection. *IEEE Int'l Workshop on Computer Science and Engineering*, 1 :534–537, 2009.

[53] R. Szeliski. *Computer Vision : Algorithms And Applications*, pages 575–643. Springer, 2011.

[54] Y. Tian, L. Brown, A. Hampapur, M. Lu, A Senior and C. Shu. IBM smart surveillance system (s3) : Event based video surveillance system with an open and extensible framework. *Machine Vision and Applications*, 19(5–6) :315–327, 2008.

[55] P. Viola and M.J. Jones. Robust real-time face detection. *Int'l Journal of Computer Vision*, 57(2) :137–154, 2004.

[56] S. Wang and Y. Wang. Simultaneous object recognition and localization in image collections. *IEEE Int'l Conference on Advanced Video and Signal Based Surveillance*, pages 497–504, 2010.

[57] J. Yagnik and M. Zhao. Automatic large scale video object recognition. `http://www.google.com/patents/US8254699/`, 2012.

[58] J. Yang, Y.G. Jiang, A.G. Hauptmann and C.W. Ngo. Evaluating bag-of-visual-words representations in scene classification. *ACM Int'l Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.

[59] M. M. Yeung, B. Yeo, W. H. Wolf and B. Liu. Video browsing using clustering and scene transitions on compressed sequences. *Multimedia Computing and Networking*, SPIE 2417 :399–413, 1995.

[60] H.J. Zhang, C.Y.Low, S.W. Smoliar and J.H. Wu. Video parsing, retrieval and browsing : An integrated and content-based solution. *ACM Int'l Conference on Multimedia*, pages 15–24, 1995.

[61] L. Zhang, S.Z. Li, X. Yuan and S. Xiang. Real-time object classification in video surveillance based on appearance learning. *IEEE Conference on Computer Vision*

and Pattern Recognition, pages 1–8, 2007.

[62] H. Zhanga, J.E. Frittsb and S.A. Goldmana. Image segmentation evaluation : A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2) :260–280, 2008.

[63] J. Zhao, S. Zhou, J. Sun and Z. Li. Point pattern matching using relative shape context and relaxation labeling. *IEEE Int'l Conference on Advanced Computer Control*, 5 :516–520, 2010.