Détection et analyse des évènements rares par vision, dans un contexte urbain ou péri-urbain

Dieudonne Fabrice Atrevi

Centre Val de Loire UNIVERSITÉ
Spirit of the Loire Valley

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

ÉCOLE DOCTORALE MATHÉMATIQUES, INFORMATIQUE, PHYSIQUE THÉORIQUE ET INGÉNIERIE DES SYSTÈMES

LABORATOIRE : PRISME

Thèse présentée par :

Dieudonné Fabrice ATREVI

soutenue le : **[17 Juin 2019]**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Mathématique-Informatique**

Détection et analyse des événements rares par vision, dans un contexte urbain ou péri-urbain

Thèse dirigée par :

Bruno EMILE	Maître de Conférence HDR, Université d'Orléans, Laboratoire PRISME
Damien VIVET	Ingénieur-Chercheur, ISAE-SUPAERO

RAPPORTEURS :

Jenny BENOIS-PINEAU	Professeure des Universités, Université de Bordeaux, Laboratoire LaBRI
Jean-Luc DUGELAY	Professeur à EURECOM Sophia Antipolis

JURY :

Rachid HARBA	Professeur des Universités, Université d'Orléans, Laboratoire PRISME, Président du jury
Jenny BENOIS-PINEAU	Professeure des Universités, Université de Bordeaux, Laboratoire LaBRI
Jean-Luc DUGELAY	Professeur à EURECOM Sophia Antipolis
Bruno EMILE	Maitre de Conférence HDR, Université d'Orléans, Laboratoire PRISME
Damien VIVET	Ingénieur-Chercheur, ISAE-SUPAERO

Remerciements

Avant tout propos, qu'il me soit permis d'adresser mes sincères remerciements aux différentes personnes qui, d'une manière ou d'une autre, ont participé à la réalisation et à la réussite de ce travail de thèse. Durant ces trois années, vous m'avez, chacun à votre niveau, soutenu surtout pendant les moments difficiles. Veuillez trouver en ces mots, l'expression de ma profonde reconnaissance.

Plus particulièrement :

Je tiens tout d'abord à remercier Bruno EMILE et Damien VIVET qui m'ont offert l'opportunité de mener ces travaux de recherche et qui en ont assuré la direction et la supervision. Durant ces trois années, ils m'ont accordé leur confiance et m'ont enseigné à être autonome. J'ai beaucoup apprécié l'ambiance de nos discussions, le soutien ainsi que les éclairages dont j'ai pu bénéficier. Votre disponibilité presque instantannée, que ce soit sur des questions administratives, scientifiques à travers vos différentes propositions, les relectures et corrections des différentes publications, a été un facteur essentiel de la résussite de ce travail. Veuillez recevoir ici ma profonde gratitude.

Je remercie également les rapporteurs de cette thèse, la professeure Jenny BENOIS-PINEAU et le professeur Jean-Luc DUGELAY, pour leur disponibilité mais également pour les suggestions et les corrections qui ont contribué à parfaire ce document. Je n'oublierai pas les autres membres du jury, le professeur Rachid HARBA qui a également pris de son temps pour juger ce travail.

Je suis reconnaissant envers la région centre val de Loire pour le financement de ces travaux. Un merci également à toutes les personnes qui ont participé à l'élaboration et ainsi qu'à la réalisation de ce projet avec en tête M Yves PARMANTIER.

Les travaux de cette thèse ont été réalisés au sein de l'équipe Image-Vision du laboratoire PRISME sur le site de Polytech Galilée à l'université d'Orléans. Je passe par ce canal pour remercier toutes les personnes qui y travaillent pour leurs aides et pour leur bonne humeur. Un spécial merci aux sécretaires Laure SPINA et Sylvie PLESSARD pour leur formidable accompagnement lors des différentes missions que j'ai pu effectuer.

Plus que des collègues, nous formions une famille "La team Galilée". A tous les collègues du site de Galilée : Kaouther, Asma, Omar, Dian, Yasmine, Antonio, Teddy, Khadija, Meina, Rania, Koubouratou et Asma Rebhi, je dis merci pour tous ces moments de gaieté que nous avons vécu ensemble. Je vous souhaite beaucoup de courage et une très bonne carrière. Mes remerciements vont aussi à l'endroit de tous les stagiaires qui ont pris par l'équipe pendant mes trois années spécialement Hamza, Rubens, Yuyuan et Charbel, mes quatre stagiaires, dont les travaux ont

contriubé à l'accomplissement de cette thèse et à l'achèvement du projet Lumineux.

Je n'oublierai pas mes amis de l'IFI : Félix, Gildas, Landy, Lionel, Der, Flore, Luyen, Paul et les autres qui, malgré la fin de nos études à Hanoï, ont su maintenir ce lien amical et fraternel qui existe entre nous. Nos partages d'expériences, nos call-conf ont été d'un grand reconfort surtout dans les moments difficiles. Merci pour ces moments.

A tous mes amis du monde entier en particulier Carole, Nhung, Loukmane, Gaël, Edgard, Serge, Mariam, Delali, Ménélick, Prudence, Naurice, merci pour le soutien indéfectible et quasi-quotidien.

A Delali, Ménélick, Carole et Sorel, un merci particulier pour la relecture des chapitres du manuscrit.

Je tiens enfin à remercier du plus profond de mon coeur mes géniteurs : Dr Nicolas ATREVI et Hélène AGADAME, mes frères et soeurs : Sorel, Mariane, Romaric, Arielle, Olivier et tous mes proches qui, malgré la distance, ont toujours été présents à mes côtés. Veuillez trouver à travers ce document ma reconnaissance pour vos nombreux efforts. Spécial dédicace à vous.

Résumé

La détection et la reconnaissance d'événements rares sont aujourd'hui des problématiques majeures pour la gestion des lieux publics ou la compréhension fine d'une scène. C'est notamment le cas dans le cadre de la gestion de l'éclairage public, qui reste aujourd'hui un enjeu essentiel pour la protection de l'environnement ainsi que pour les économies d'énergie des communautés urbaines. Pour atteindre ce but et favoriser le développement des villes intelligentes sans mettre en péril la sécurité des personnes et des biens, des techniques d'analyse de scènes et plus particulièrement de détection d'événements sont nécessaires pour affiner les prises de décision.

L'objectif principal de cette thèse est le développement de méthodes complètes de détection d'événements rares. Les travaux effectués portent sur deux thématiques complémentaires. La première thématique est consacrée à l'étude des descripteurs de formes pour la détection et la reconnaissance d'éléments d'une scène. D'une part, la robustesse de certains descripteurs face à différentes conditions de luminosité a été étudiée. D'autre part, les moments géométriques ont été comparés à travers une application d'estimation de pose humaine 3D à partir d'image 2D. De cette étude, nous avons montré qu'avec une application de recherche de formes, les moments géométriques permettent d'estimer la pose d'une personne à travers une recherche exhaustive dans une base d'apprentissage de poses connues. Cette application peut être utilisée dans un système de reconnaissance d'actions pour une analyse plus fine des événements détectés. En ce qui concerne la deuxième thématique qui est relative à la détection et la localisation des événements rares, trois contributions sont présentées. La première contribution concerne l'élaboration d'une méthode d'analyse globale de scène pour la détection des événements liés aux mouvements de foule. Dans cette approche, la modélisation globale de la scène est faite en nous basant sur des points d'intérêt filtrés à partir de la carte de saillance de la scène. Les caractéristiques exploitées sont l'histogramme des orientations du flot optique et un ensemble de descripteur de formes étudié dans la première partie. L'algorithme LDA (Latent Dirichlet Allocation) est utilisé pour la création des modèles d'événements à partir d'une représentation en document visuel de séquences d'images (clip vidéo). La deuxième contribution consiste en l'élaboration d'une méthode de détection de mouvements saillants ou dominants dans une vidéo. La méthode, totalement non supervisée, s'appuie sur les propriétés de la transformée en cosinus discrète pour analyser les informations du flot optique de la scène afin de mettre en évidence les mouvements dominants. La modélisation locale pour la détection et la localisation des événements est au coeur de la dernière contribution de cette thèse. La méthode se base sur les scores de saillance des mouvements et de l'algorithme SVM dans sa version "one class" pour créer le modèle d'événements. Les méthodes ont été évaluées sur différentes bases publiques et les résultats obtenus sont prometteurs.

Abstract

The efficient management of public lighting in order to protect the environment and to save energy, without compromising the safety of people and property, is a mainstay in the development of smart cities. The work carried out during this thesis is part of the LUMINEUX project, which aims to provide technological solutions for intelligent public lighting management by taking into account events occurring around the light pole.

The main objective of this thesis is the development of complete methods for rare events detection. The works can be summarized in two parts. The first part is devoted to the study of shapes descriptors of the state of the art. On the one hand, the robustness of some descriptors to varying light conditions was studied. On the other hand, the ability of geometric moments to describe the human shape was also studied through a 3D human pose estimation application based on 2D images. From this study, we have shown that through a shape retrieval application, geometric moments can be used to estimate a human pose through an exhaustive search in a pose database. This kind of application can be used in human actions recognition system which may be a final step of an event analysis system. In the second part of this report, three main contributions to rare event detection are presented. The first contribution concerns the development of a global scene analysis method for crowd event detection. In this method, global scene modeling is done based on spatiotemporal interest points filtered from the saliency map of the scene. The characteristics used are the histogram of the optical flow orientations and a set of shapes descriptors studied in the first part. The Latent Dirichlet Allocation algorithm is used to create event models by using a visual document representation of image sequences (video clip). The second contribution is the development of a method for salient motions detection in video. This method is totally unsupervised and relies on the properties of the discrete cosine transform to explore the optical flow information of the scene. Local modeling for events detection and localization is at the core of the latest contribution of this thesis. The method is based on the saliency score of movements and one class SVM algorithm to create the events model. The methods have been tested on different public database and the results obtained are promising.

Sommaire

Liste des tableaux

Liste des figures

Liste des sigles et Abréviations

ACP : Analyse en Composante Principale

AUC : Area Under Curve

BGMMAI : Bayesian Generative Model based on Motion and Appearance Information

CHOP : Color Histogram of Oriented Phase

CNN : Convolutional Neural Network

DCT : Discrete Cosine Transform

DFT : Discrete Fourier Transform

DPM : Deformable Part Models

EER : Equal Error Rate

EM : Expectation-Maximization

ESS : Efficient Sub-window Search

HOG : Histogram of Oriented Gradient

KNFST : Kernel Null Foley-Sammon Transform

LBP : Local Binary Pattern

LDA : Latent Dirichlet Allocation

LSTM : Long Short-Term Memory

MDT : Mixture Dynamic Texture

MHOF : Multi-scale Histogram of Optical Flot

MPPCA : Mixture of Probabilistic Principal Component Analysers

MRF : Markov Random Field

NMS : Non Maximum Suppression

RD : Rate Detection

ReLU : Rectified Linear Unit

RNN : Recurrent Neural Network

ROC : Receiver Operating Characteristics

SIFT : Scale-Invariant Feature Transform

SRC : Sparse Reconstruction Cost

STIP : Spatio-temporal Interest Point

SURF : Speeded Up Robust Features

SVM : Séparateurs à Vastes Marges

TFP : Taux de Faux Positif

TVP : Taux de Vrai Positif

Introduction

Cadre des travaux de la thèse

Les travaux de recherche présentés dans ce rapport de thèse ont été réalisés dans le cadre du projet LUMINEUX, financé par la région « Centre Val de Loire ». Le but principal du projet est d'assurer une économie énergétique et une protection environnementale aux villes tout en assurant la sécurité des personnes et des biens. L'économie énergétique et la protection environnementale à travers la gestion efficace et intelligente de l'éclairage public sont devenues des enjeux majeurs dans l'atteinte de l'objectif des villes intelligentes ou villes du futur.

En effet, selon l'Agence de L'Environnement et de la Maîtrise de L'Energie (ADEME) [1], l'éclairage public représente 41% des consommations d'électricité et 37% des factures d'électricité des collectivités territoriales. Ce coût énorme lié à la consommation d'électricité combiné avec les effets nuisibles des éclairages publics sur l'environnement, sont quelques raisons principales qui poussent certains décideurs à éteindre l'éclairage public à partir d'une heure donnée. La réglementation de ces décisions, qui peuvent être perçues comme radicales, est assurée à travers la loi portant « engagement national pour l'environnement » dite « Grenelle 2 » du 12 juillet 2010, qui en son article 173 stipule : « Pour prévenir ou limiter les dangers ou troubles excessifs aux personnes et à l'environnement causés par les émissions de lumière artificielle et limiter les consommations d'énergie, des prescriptions peuvent être imposées, pour réduire ces émissions, aux exploitants ou utilisateurs de certaines installations lumineuses, sans compromettre les objectifs de sécurité publique et de défense nationale ainsi que de sûreté des installations et ouvrages sensibles ». Ainsi, toute collectivité, visant à instaurer une nouvelle politique de gestion de l'éclairage public, doit s'assurer que ces mesures n'impactent pas la sécurité publique. Des études ont montré que l'éclairage public dans les zones urbaines participe à une diminution de 30% à 40% des accidents de la circulation et de 50% des effractions, vols et actes de vandalisme. Le recours à l'extinction complète de l'éclairage ne peut être une solution efficicace dans l'atteinte des objectifs sus-mentionnés. Une gestion intelligente de ces installations s'impose et représente un défi tant pour les décideurs que pour les constructeurs de luminaires.

Les nouvelles technologies sont mises à contribution pour mettre en place des solutions innovantes et intelligentes dans la gestion des villes. L'introduction des dalles à base de LED dans les luminaires permet d'envisager à court et moyen terme des systèmes d'éclairage intelligents afin de nuancer la puissance de l'éclairage en fonction de la scène observée. D'après une étude

1. `https://www.ademe.fr/collectivites-secteur-public/patrimoine-communes-comment-passer-a-laction/eclairage-public-gisement-deconomies-denergie`, consulté le 04/07/2018

menée par l'agence de développement pour la province de Liège, une estimation d'une réduction de 65% de la consommation est possible grâce à un système de gestion intelligente de l'éclairage (15% de réduction supplémentaire grâce à l'utilisation de LED). Le projet Lumineux vise à proposer des solutions innovantes pour atteindre le triple objectif : économie d'énergie, protection environnementale et sécurité des personnes et des biens.

Motivations et Objectifs de la thèse

La détection d'événements rares fait partie des axes dynamiques de la recherche en vision par ordinateur et plus particulièrement en analyse de scène. Elle attaque une problèmatique majeure de notre société qui est la sécurité des personnes et des biens par ces temps caractérisés par des attentats dans les espaces publics de forte concentration humaine. Deux approches se distinguent pour ce type d'analyse : la première est orientée vers la détection des objets et la seconde, plus récente, s'appuie sur des connaissances apprises au « fil du temps ». La première classe de méthodes se base donc principalement sur une approche supervisée qui consiste à apprendre hors-ligne les actions ou/et les objets recherchés et à les reconnaître dans la scène observée. Ces méthodes s'appuient sur une architecture en cascade où les objets sont tout d'abord détectés, suivis puis classifiés. Ensuite, le chemin parcouru par les différents objets détectés est analysé puis confronté à un modèle pour détecter les situations suspectes. Parmi les travaux utilisant cette approche, on peut citer les exemples suivants : la reconnaissance d'actions [1], la détection d'objets abandonnés [2][3] ou le comptage de véhicules [4]. On parle dans la littérature des méthodes de reconnaissance ou de classification d'actions. Les bases de données fournies dans littérature, contiennent des groupes d'actions tels que : marche, jogging, courrir, tomber, etc. Chaque groupe contient plusieurs vidéos dans lesquelles les mêmes actions sont effectuées mais avec différents acteurs. Les difficultés résident alors dans le choix d'un apprentissage représentatif et varié permettant de reconnaître les différentes situations dans des contextes distincts. Cette catégorie d'approches est parfaitement compatible avec des applications pour lesquelles une liste exhaustive des situations à risque sont bien identifiées. Néanmoins, cela représente une des faiblesses de cette catégorie d'approches, pour son utilisation dans des conditions réelles non toujours maîtrisées.

La seconde catégorie propose un apprentissage au « fil du temps » de la scène observée. Ces méthodes permettent de construire un modèle « normal » des événements qui se produisent fréquemment. L'objectif est alors la détection de valeurs aberrantes (ou « outliers » courramment utilisé dans la littérature), qui consiste à rejeter à partir d'un seuil, toute donnée n'ayant pas de forte similarité avec le jeu d'apprentissage. Ainsi, à travers une représentation spatio-temporelle de la scène, les événements « rares » peuvent être détectés [5]. Dans [6], Adam et al. proposent une solution basée sur le flot optique où, après un apprentissage statistique d'évolution des vecteurs, les activités anormales sont identifiées grâce aux mesures de déviations de ces vecteurs. Certaines approches proposées sont très prometteuses (voiture faisant demi-tour à un carrefour, personne déposant un colis), mais restent pour le moment difficiles à mettre en œuvre dans des contextes moins structurés (mouvements désordonnés), comme par exemple dans le cas de la détection de

chute.

Les travaux réalisés dans le cadre de cette thèse se situent dans la deuxième catégorie. Ce positionnement est dû à la fois au cadre de réalisation de la thèse car l'identification de toutes les situations potentiellement anormales est impossible, mais aussi dû à l'avancement de l'état de l'art sur la thématique. Le niveau de détection dans la scène dépend du niveau de traitement. On considère généralement deux niveaux de traitement : le traitement global et celui local. Le premier niveau de traitement vise à faire une détection précoce en alertant sur tout soupçon de présence de situations non connues (rares) dans une séquence d'images. La scène est considérée dans sa globalité sans recherche de localisation des différents événements. Même si l'extraction des caractéristiques peut se faire dans différentes zones de l'image, le traitement au niveau globale vise à faire une modélisation des événements en utilisant une représentation compacte de séquences d'images (clip), c'est à dire un vecteur unique de descripteurs par clip. Le second niveau de traitement va plus en profondeur dans la modélisation des événements en proposant une modélisation locale. Dans ce cas, la recherche se fait zone par zone en comparant les informations de chaque zone au modèle d'événements appris. Le premier niveau de traitement peut servir dans une couche supérieure d'un système d'alerte et servir de déclencheur pour le second niveau de traitement qui ira localiser l'événement et si possible en déterminer la nature. Nous avons proposé au cours de nos travaux, des méthodes pour les deux niveaux sus-mentionnés.

Les objectifs visés par cette thèse peuvent se résumer ainsi :

- le développement de méthodes pour la détection et la localisation d'événements rares dans une scène naturelle : il s'agit de développer de nouvelles approches de modélisation d'événements qui prennent en compte plusieurs facteurs qui peuvent influencer la réussite des détections ;

- l'étude et la proposition de descripteurs robustes et adaptés à la modélisation d'événements pour différentes situations : les descripteurs sont un maillon essentiel dans la chaîne de traitement et de modélisation car leur qualité influence beaucoup sur la précison de la détection. Chaque descripteur vise à représenter d'une manière donnée l'information dans la scène. Le but du travail sera de trouver des descripteurs pertinents dont l'utilisation par la suite pour la modélisation permettra de couvrir le maximum de cas d'événements ;

- l'étude d'algorithmes d'apprentissage adaptés à la modélisation d'événements : le type de problème que nous traitons appartient à la catégorie de la classification "one class". Cela suppose que dans les jeux d'apprentissage, les données représentent à peu près les mêmes événements. Des différences énormes peuvent exister au sein d'un même groupe d'événements (différence intra-événement). Le but du travail ici, sera de trouver des algorithmes d'apprentissage qui pourront efficacement produire des modèles qui englobent le plus de situations présentes dans le jeu d'apprentissage tout en permettant la détection des événements rares ;

Contributions scientifiques et plan de rédaction

Les travaux réalisés ont eu pour but de proposer de nouvelles approches pour la détection et la localisation d'événements rares ou anormaux dans une vidéo. Les contributions scientifiques

apportées par cette thèse se situent à différents niveaux. Ainsi, l'extraction de caractéristiques pertinentes pour l'apprentissage des événements récurrents est une étape très importante dans la chaîne de détection. Nous avons étudié plusieurs descripteurs de la littérature pour sélectionner les plus adaptés à notre contexte. Nous avons également étudié d'autres descripteurs connus dans la littérature et utilisés pour diverses applications de vision par ordinateur. Le chapitre 1 est le condensé des différentes études menées sur ces descripteurs.

En ce qui concerne les travaux qui sont au coeur de cette thèse, deux méthodes ont été proposées. Dans un premier temps, le chapitre 2 est consacré à un état de l'art des méthodes récentes proposées sur la thématique. Dans un second temps, les deux méthodes proposées seront abordées dans les chapitres suivants. Ainsi, la première méthode qui vise un traitement au niveau global des informations pour une détection de premier niveau est présentée dans le chapitre 3. Cette méthode est basée sur deux types de descripteurs à savoir celui du mouvement et celui de l'apparence. Nous avons eu recours à certains descripteurs présentés au niveau du chaptire 1 pour l'extraction des informations utiles pour notre méthode. La modélisation des événements est basée sur l'utilisation de l'algorithme LDA (Latent Dirichlet Allocation) utilisé dans le domaine du traitement de texte. Une évaluation de la méthode sur la base publique UMN consacrée à la détection de mouvement de panique de foule, a montré que notre méthode est compétitive par rapport à celles de l'état de l'art. La valorisation du travail est faite à travers différentes publications scientifiques [7, 8]. La deuxième méthode qui vise la détection et la localisation des événements dans la scène a été présentée dans le chapitre 4. Il s'agit d'un traitement de second niveau visant une modélisation locale des informations extraites localement de la scène. Dans ce chapitre, nous avons présenté dans un premier temps une méthode de détection de mouvements saillants. Le mouvement saillant par définition est un mouvement irrégulier au regard de tous les mouvements en cours dans une scène. Notre méthode est basée sur les propriétés de la transformée en cosinus discrète à travers la reconstruction d'images à partir des signes des coefficients de sa transformée. La détection des mouvements saillants est une étape de sélection de régions potentielles contenant des événements rares. Par la suite nous avons proposé une approche de modélisation et de détection basée sur les scores des mouvements saillants précédemment identifiés dans la scène. Les performances issues des expériences menées montrent que notre approche est prometteuse au regard de l'état de l'art. Cette dernière partie a été valorisée à travers différentes publications scientifiques [9, 10].

Des travaux annexes ont porté sur l'étude des moments géométriques pour une application d'estimation de pose humaine 3D à partir d'images 2D. Les moments géométriques servent de descripteurs pour le matching entre différentes formes enregistrées dans une base d'images. La contribution se trouve au niveau de l'étude comparative et l'application à un problème concret. Le but est d'utiliser in fine cette méthode de reconnaissance d'action pour une éventuelle caractérisation des événements détectés. Les contributions dans cette partie ont été valorisées par différentes publications scientifiques [11, 12].

Chapitre 1

Détection d'objets basée sur les descripteurs de formes

> L'imagination est plus importante que la connaissance. Car la connaissance est limitée, tandis que l'imagination englobe le monde entier, stimule le progrès, suscite l'évolution.
>
> Albert Einstein

Sommaire

1.1 Introduction

Le cadre des travaux de cette thèse, comme mentionné dans l'introduction de ce document, concerne la gestion intelligente de l'éclairage public en fonction des objets présents sous les luminaires et des événements qui s'y déroulent. Nous nous intéressons alors à la détection d'activités, d'événements et d'objets dans les vidéos issus des caméras positionnées au niveau des dits luminaires. Pour ce faire, il faut être capable de détecter les caractéristiques de chaque objet ou événement ; caractéristiques qui passent par l'apparence visuelle des scènes. L'extraction de ces caractéristiques reposent sur des descripteurs de diverses natures : de forme, de mouvement, etc. Les conditions de luminosité de l'application nous amènent à la recherche de descripteurs invariants pour des détections en pleine journée ou dans la nuit profonde avec une variation fréquente de l'intensité lumineuse.

Ces descripteurs, au-delà de leur utilisation pour la détection de piétons, voitures, cycles, etc., peuvent servir dans les travaux sur l'élaboration de méthodes de détection d'événements rares. Ainsi, ce chapitre sera consacré à l'étude de descripteurs de formes et de méthodes existants dans la littérature pour la détection de piétons, voitures etc. Le but est de trouver des descripteurs robustes aux changements de luminosité dans la scène. En effet, à défaut d'avoir à entraîner des modèles pour la reconnaissance d'objets dans le visible (en pleine journée) ou dans le nocturne, il serait plus judicieux d'avoir un seul modèle appris sur des données obtenues dans toutes les conditions de luminosité et de pouvoir procéder à des détections dans n'importe quelle condition. La plupart des bases de données d'objets ou de personnes utilisées dans la littérature pour l'apprentissage et le test contiennent des images de jour. L'évaluation des descripteurs sera faite dans le cadre de la détection de personnes. A travers notre étude, nous présenterons les descripteurs et algorithmes d'apprentissage les plus pertinents de la littérature ainsi que les méthodes les plus récentes. Nous étudierons l'influence de la combinaison de certains de ces descripteurs afin de trouver les plus adaptés pour résoudre la problématique posée. L'étude des moments géométriques pour la description de forme a également été menée et est présentée en annexe A. Notre contribution dans le présent chapitre se situe au niveau de l'évaluation de l'efficacité de quelques descripteurs de formes dans des conditions de détection difficiles et aussi de l'influence de leur différentes combinaisons sur les performances de la détection.

Les descripteurs de formes sont classés en deux catégories à savoir : les descripteurs locaux et globaux. Par descripteurs locaux, on parle de descripteurs extraits dans des zones éparses d'une image, généralement autour de points d'intérêt et qui décrivent localement ce dernier en utilisant les informations de son environnement immédiat. D'un autre côté, les descripteurs globaux sont utilisés pour décrire une image entière.

1.2 Les descripteurs locaux

Dans cette section, nous allons présenter quelques descripteurs locaux populaires dans la littérature : le descripteur Haar-like, le descripteur SIFT et le descripteur SURF.

1.2.1 Le descripteur Haar-like

Introduit par Viola et Jones [13] en 2001 pour la détection de visage, le descripteur Haar-like est inspiré des ondelettes de Haar utilisées en traitement du signal. Il s'agit d'un descripteur simple et rapide basé sur le calcul de la différence de la somme de l'intensité des pixels de régions adjacentes (région blanche et région noire). Au-delà d'un certain seuil prédéfini, le résultat de la détection du motif recherché est positif.

$$\sum(regionBlanche) - \sum(regionNoire) > \theta$$

L'arrangement spatial des régions adjacentes, correspond à un type particulier de motif (contour, ligne, etc.) à détecter (voir Figure 1.1). Ces arrangements peuvent être vu comme des filtres. Les deux régions adjacentes forment ensemble la région de détection. On peut ainsi dire que le résultat est la réponse de la région pour un filtre précis. Prenons deux régions adjacentes $regionBlanche$ et $regionNoire$. Il est démontré et admis qu'un contour dans une image est la frontière entre deux régions adjacentes avec une grande différence d'intensité. Le motif (a) de la figure 1.1 permet, en calculant la différence entre la somme des intensités de pixel de la région blanche et de la région noire, de savoir si un contour se trouve dans la zone considérée.

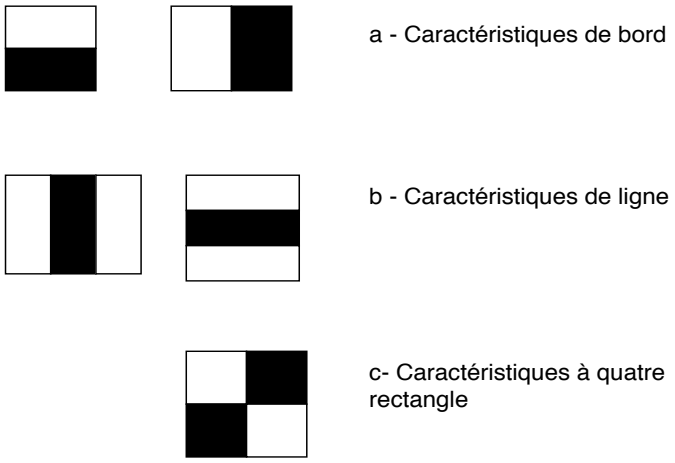


FIGURE 1.1 – Exemple de motifs de bases du descripteur Haar-like. Source : OpenCV doc `https://docs.opencv.org/3.4.1/d7/d8b/tutorial_py_face_detection.html`

Pour arriver à détecter un objet, il est procédé au calcul d'un grand nombre de réponses pour plusieurs régions de l'image avec différents types de filtres de tailles variées. Ce calcul peut rendre le descripteur lent pour certains types d'objets à détecter. Pour éviter le calcul à répétition de la somme dans chaque région de l'image, la notion d'image intégrale a été introduite. A partir de l'image intégrale, on peut obtenir la somme des pixels dans toutes les zones de l'image. Dans une image intégrale, la valeur d'un pixel à la position (x, y) est la somme des pixels situés au-dessus et à gauche de cette position. Ainsi, dans la figure 1.2, les points 1, 2, 3, 4 correspondent

respectivement aux sommes des pixels contenus dans les zones A, $(A+B)$, $(A+C)$ et $(A+B+C+D)$. Pour obtenir la somme des valeurs des pixels de la zone D uniquement, il faut faire $4-2-3+1$. Le descripteur Haar-like a été utilisé dans une méthode de détection de piétons

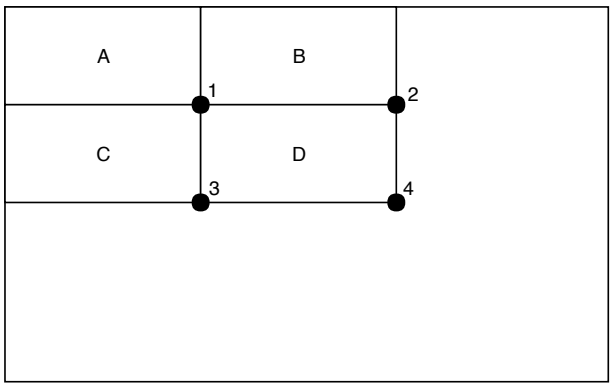


FIGURE 1.2 – Exemple d'image intégrale

par Papageorgiou et al. [14]. Sa performance bien qu'étant limitée, était prometeuse et peut permettre une analyse rapide de zones d'images dans une application de recherche de motifs simples.

1.2.2 Les détecteurs-descripteurs SIFT et SURF

Les descripteurs SIFT (Scale-Invariant Feature Transform) et SURF (Speed-Up Robust Features) sont des descripteurs utilisés conjointement avec un détecteur de points d'intérêt dans une image (des détails sur les points d'intérêt sont donnés dans le chapitre 3). Les points d'intérêt détectés avec différentes approches, sont décrits par le descripteur qui est utilisé pour l'extraction d'un ensemble d'informations dans l'environnement de chaque point d'intérêt.

Le descripteur SIFT, proposé par Lowe et al. [15], est une représentation, sous forme d'histogramme, des informations de gradients dans l'entourage des points d'intérêt. En effet, comme le montre la Figure 1.3, les orientations et amplitudes des vecteurs de gradients sont caculés pour chaque pixel (voir à gauche sur l'image) dans une région, idéalement de taille 16x16 (mais 8x8 sur l'image), centrée sur chaque point d'intérêt. Les gradients sont pondérés par une fenêtre gaussienne représentée par le cercle en bleu. Il s'ensuit alors l'étape de regroupement des informations, bloc par bloc de taille 4x4 (comme sur la figure), sous forme d'histogramme à 8 orientations, en additionnant les amplitudes des vecteurs ayant presque les mêmes directions. Les histogrammes des 16 blocs sont par la suite concatenés pour donner le descripteur SIFT, de taille 128 dimensions (16×8), du point d'intérêt. Le contenu d'une image est alors décrit par l'ensemble des vecteurs SIFT des points d'intérêt qui s'y trouvent. Le descripteur est invariant à l'orientation et à la résolution de l'image, et est peu sensible à son exposition, à sa netteté ainsi qu'au point de vue 3D, ce qui constitue un avantage par rapport à d'autres descripteurs. Cette robustesse vaudra au descripteur d'être utilisé pour plusieurs tâches dans des applications de recherche d'éléments similaires dans des images, de reconnaissance d'objets, de détection de personnes, de cartographie, de navigation, de suivi d'objets, etc. Sa lourdeur en

terme de temps de calcul est son talon d'Achille. Il est important de noter que la méthode SIFT utilise un ensemble de techniques pour élaborer son propre algorithme de détection de points d'intérêt multi-échelle. A chaque point détecté est assigné une position 2D, un facteur d'échelle et une orientation (orientation dominante dans son entourage).

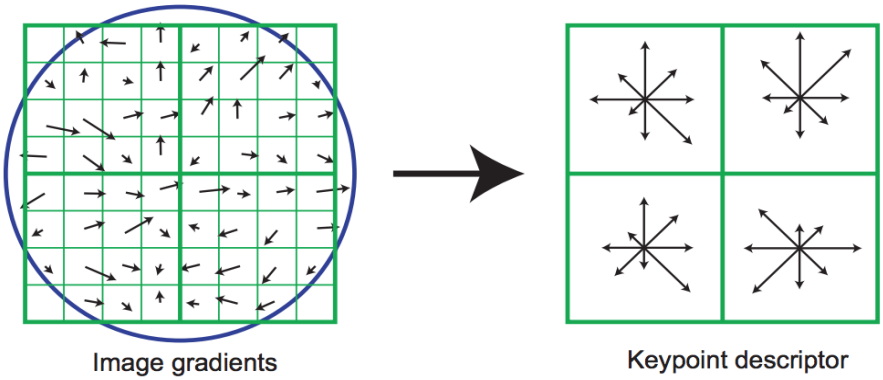


FIGURE 1.3 – Structure du descripteur SIFT [15]

Pour rémédier aux problèmes liés au temps de calcul et à certaines transformations, le descripteur SURF a été introduit par Bay et al. [16] dans les années 2006. Il s'inspire en partie du descripteur SIFT mais est fondé sur la somme de réponses d'ondelettes de Haar 2D calculée à l'aide d'images intégrales. Ses points d'intérêt sont détectés dans des structures de type blog où le déterminant de la matrice hessienne est maximum. Les points sont également invariants à l'échelle et à la rotation et sont rapides à calculer. Comme le descripteur SIFT, à chaque point est associé l'échelle de détection et son orientation. A l'opposé du descripteur SIFT, SURF décrit la distribution des réponses des ondelettes de Haar dans l'entourage de chaque point d'intérêt par des vecteurs de caractéristiques de 64 dimensions. Pour ce faire, des fenêtres de tailles $20s$ (s étant le facteur d'échelle associé au point) sont positionnées, centrées sur chaque point et orientées selon l'orientation du point. Ces fenêtres sont subdivisées en 16 sous-régions (4x4) et les réponses des filtres de Haar horizontal et vertical (présentés ci-dessus) sont calculées dans chaque sous-région (en utilisant des fenêtres espacées de taille 5x5). Les réponses des filtres sont lissées avec un filtre gaussien de variance $\sigma = 3.3s$ centré sur le point d'intérêt. Toutes les réponses des filtres sont additionnées dans chaque sous-région. Soit d_x la réponse du filtre horizontal et d_y celle du filtre vertical. Le vecteur de descripteur d'une sous-région est constituée de : $v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$. L'ensemble des 16 vecteurs de taille 4 sont concatenés pour donner un vecteur SURF de taille 64 qui représente le point d'intérêt. On peut voir sur la figure 1.4 différentes valeurs des réponses des ondelettes pour différents types de région. La variation observée au niveau de l'image du milieu et de droite est bien traduite par les composantes $\sum d_x$ et $\sum |d_x|$ de v. Dans leur papier, les auteurs ont conclu que le descripteur en plus d'être plus rapide que SIFT est plus performant dans plusieurs applications. L'utilisation de ces deux caractéristiques se fait à travers un processus de "sac de mots visuels" (Bag Of Visual Word en anglais). Ce processus permet de représenter les images sous forme de documents qui contiennent des mots visuels. Les mots sont issus d'algorithmes de regroupement de données

FIGURE 1.4 – Exemple du vecteurs de SURF pour différents types de régions [16]

telque le KMeans, appliqués à un ensemble de vecteurs de caractéristiques SIFT ou SURF. La reconnaissance d'objet basée sur un tel processus se fait en construisant des modèles d'objets à partir des vecteurs d'occurrence des mots visuels qui représentent l'image et non directement à partir des vecteurs SIFT ou SURF.

1.3 Les descripteurs globaux

En plus des descripteurs locaux, qui décrivent des zones précises d'une image, généralement autour de points d'intérêt, d'autres descripteurs décrivent globalement une image ou fenêtre d'une image. Il s'agit des descripteurs globaux. Dans le cas des descripteurs locaux tel que SIFT, un objet est décrit par un ensemble de vecteurs de descripteurs de points d'intérêt. Les descripteurs globaux s'affranchissent des points d'intérêt et permettent de décrire les objets dans leur globalité. Pour les tâches de détection de personnes par exemple, cette catégorie de descripteurs a eu un franc succès à l'image du descripteur HOG (Histogram of Oriented Gradient).

1.3.1 Le descripteur HOG

Le descripteur HOG présente quelques similarités avec le descripteur SIFT en ce sens qu'il décrit les images à travers la distribution des orientations des gradients (voir Fig 1.5). Inventé par Dalal et al. [17] pour, au départ, la détection de personnes dans une image, le descripteur HOG diffère du descripteur SIFT par l'utilisation d'une grille dense (zones régulièrement réparties) dans l'étape de calcul des histogrammes. Dans sa version originale, destinée à la détection de personnes, les images d'entrée sont redimenssionées à la taille 64×128. Le gradient est calculé, avec des filtres de Sobel, pour tous les pixels de l'image ainsi redimensionnée et les informations d'orientation et de magnitude sont extraites. L'image est subdivisée en des cellules de taille 8×8 sur une grille dense. Les histogrammes des orientations de 9 orientations à l'intérieur de chaque cellules sont calculés en pondérant le vote de bins correspondant à l'orientation par l'amplitude associée. De ce fait, les histogrammes sont sensibles au changement de luminosité dans l'image. Pour corriger le problème, les auteurs proposent une normalisation L2 des histogrammes des

cellules, dans des blocs de taille 16×16 px (ou 2×2 cellules) collectés sur toute la fenêtre avec un décalage de 8 pixels chaque fois. Les quatre histogrammes de chaque bloc sont concatenés avant la normalisation pour donner un vecteur de 36 dimmensions par bloc. Pour une image d'entrée, on peut recenser 105 blocs. Pour obtenir le descripteur HOG de l'image, tous les vecteurs des 105 blocs sont concatenés ensemble pour donner un unique vecteur de dimension 3780-D. Un modèle de la forme humaine est construit en utilisant des HOG d'images des personnes comme données positives et des images ne contenant aucun humain comme données négatives comme entrée d'un SVM à deux classes. La détection d'une personne dans une grande image est faite à travers une recherche exhaustive par la méthode de la fenêtre glissante.

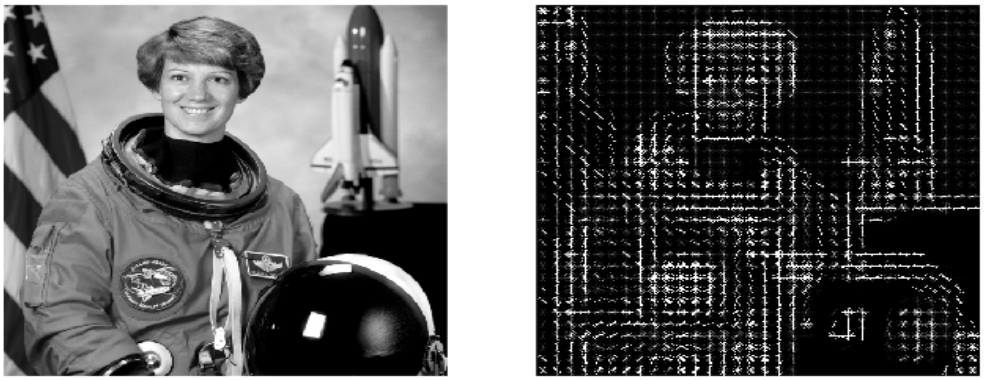


FIGURE 1.5 – Visualisation des vecteurs de gradients

Le schéma d'extraction décrit plus haut, peut être résumé avec le diagramme de l'image 1.6.

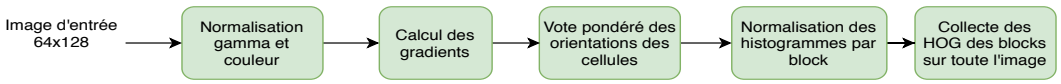


FIGURE 1.6 – Différentes étapes d'extraction du vecteur HOG

Des améliorations ont été apportées à la version initiale du HOG pour améliorer ses performances. On peut citer le PHOG (pour Pyramid Of Histogram of Oriented Gradients) de Bosch et al. [18], qui renforce l'invariance du descripteur au facteur d'échelle. La méthode consiste à calculer pour la même image, plusieurs vecteurs HOG à différentes échelles et par la suite faire la concaténation pondérée des différents vecteurs.
La figure 1.7 montre l'extraction du vecteur PHOG pour des images avec un niveau de la pyramide égale à 2. Les poids de pondérations des vecteurs HOG de chaque niveau de la pyramide peuvent être appris ou fixer de façon empirique.

1.3.2 Le descripteur LBP

LBP (Local Binary Pattern) est un descripteur de forme, utilisé principalement pour la description des textures [19]. La méthodologie consiste à comparer le niveau de luminance d'un pixel avec ses voisins. Les P voisins du pixel central de coordonnées (x_c, y_c) sont déterminés en général sur un cercle par $(x_c + R.cos(2\pi p/P), y_c + R.sin(2\pi p/P))$ avec R le rayon du cercle. Une extrapolation bilinéaire permet de déterminer dans le plan image, le pixel correspondant à un

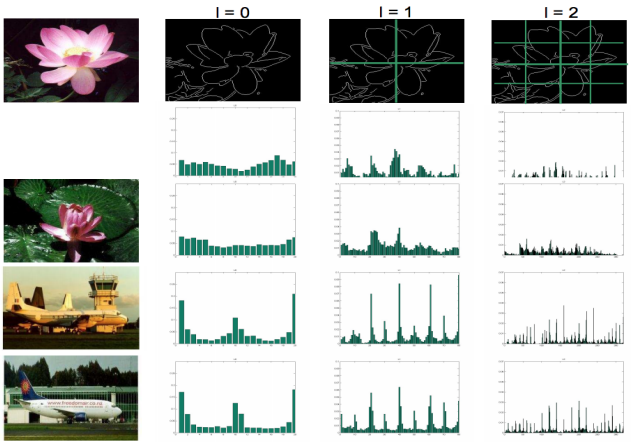


FIGURE 1.7 – Extraction de vecteurs PHOG pour différentes échelles [18]

résultat approximatif de la formule précédente.

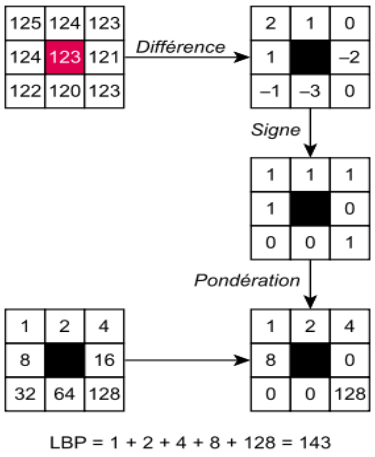


FIGURE 1.8 – Processus de calcul d'un LBP. Réalisée par Xiawi — Travail personnel, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=11747868`

L'image 1.8 décrit de façon claire les différentes étapes de calcul de la valeur LBP d'un pixel. La différence entre la valeur en niveau de gris du pixel central avec les valeurs en niveau de gris de ses voisins, représente la texture sans perte d'information. Le signe de cette différence est utilisé pour générer un masque binaire qui sera pondéré avec le facteur 2^p. La somme des valeurs pondérées du masque binaire, donne la valeur de la description locale de la texture de l'image en ce point. L'ensemble des N valeurs LBP d'une image à N pixels forme le descripteur global qui décrit la texture de l'image. Combiné avec des algorithmes d'apprentissage, le descripteur LBP a été utilisé pour de multiples tâches de reconnaissance et de classification.

Mu Yadong et al. dans [20], ont utilisé deux variantes de LBP (S-LBP et F-LBP) pour adapter la version naïve à la tâche de la détection de personne dans des images. Le S-LBP (Semantic LBP) propose de résoudre le problème de distinction de motifs sémantiquement similaires (par exemple, motifs qui diffèrent par une rotation) que rencontre le LBP traditionnel. Ils proposent d'abandonner le calcul de code LBP par la somme des valeurs pondérées du masque, qui se fait en dernière étape du LBP traditionnel, au profit d'une représentation originale. En effet, après

l'étape d'obtention du masque, ils proposent de former des arcs avec les suites successives de "1". L'angle principale et la longueur des arcs sont utilisés conjointement pour une représentation matricielle (en colonne les cases représentant les angles et en ligne les cases représentant les longueurs). Le descripteur final correspond à un vecteur unique qui est la concaténation des vecteurs colonnes de la matrice. L'image 1.9 présente un resumé du processus d'extraction du S-LBP.

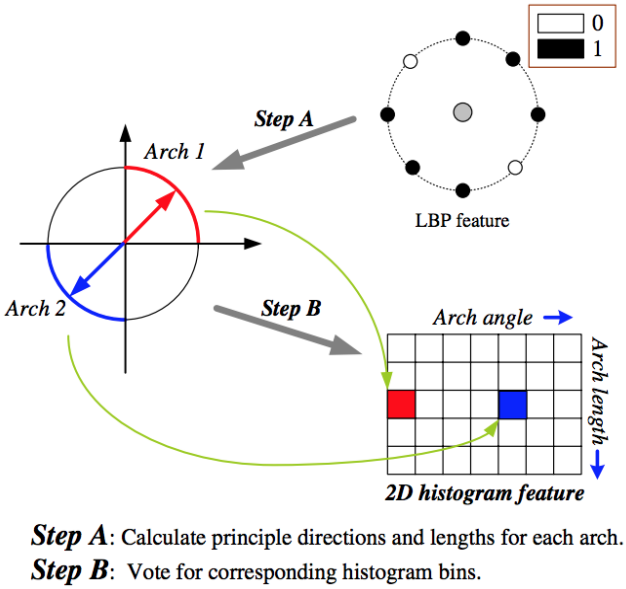


FIGURE 1.9 – Processus de calcul du S-LBP [20]

Dans la version F-LBP (Fourier LBP), pour éviter les problèmes liés au choix d'un seuil convenable dans l'étape de seuillage, ils proposent une transformation dans le domaine fréquentiel des valeurs brutes de la différence entre la valeur du pixel central et celles des pixels voisins. Soit $S = \{s(k), k = 0...P-1\}$ le résultat de la différence entre le pixel central et les voisins. A l'étape suivante, à défaut de faire un seuillage pour obtenir un masque, ils appliquent un DFT (Discrete Fourier Transform) (voir eq 1.1) sur S pour obtenir un autre vecteur $\mathcal{A} = \{a(u), u = 0...P-1\}$. Les coefficients des hautes fréquences sont éliminés pour ne conserver que les basses fréquences qui contiennent l'information de la structure locale de la saillance du motif.

$$a(u) = \frac{1}{P} \sum_{k=0}^{P-1} s(k) e^{-j2\pi uk/P} \tag{1.1}$$

1.3.3 Le descripteur CHOP

CHOP (Color Histogram of Oriented Phase) est un descripteur de formes basé sur la congruence de phase d'une image dans le domaine fréquentiel. La congruence de phase est une caractéristique invariante au contraste et à la luminosité et par conséquent utile pour la détection dans des conditions de grandes variations de luminosité (voir fig 1.10). Contraitement au gradient qui se focalise sur les zones où interviennent des changements d'intensité des pixels, la congruence de phase considère l'image comme constituée de beaucoup d'information et peu de redondances. Une analyse fréquentielle permet de mettre en évidence les contours dans une image, quelque soit

la luminosité ou le contraste. Ces contours correspondent aux points où la congruence de phase et l'énergie locale sont maximales. De plus, des études psychologiques prouvent que le système visuel de l'homme est sensible aux zones d'une image avec une information élevée.

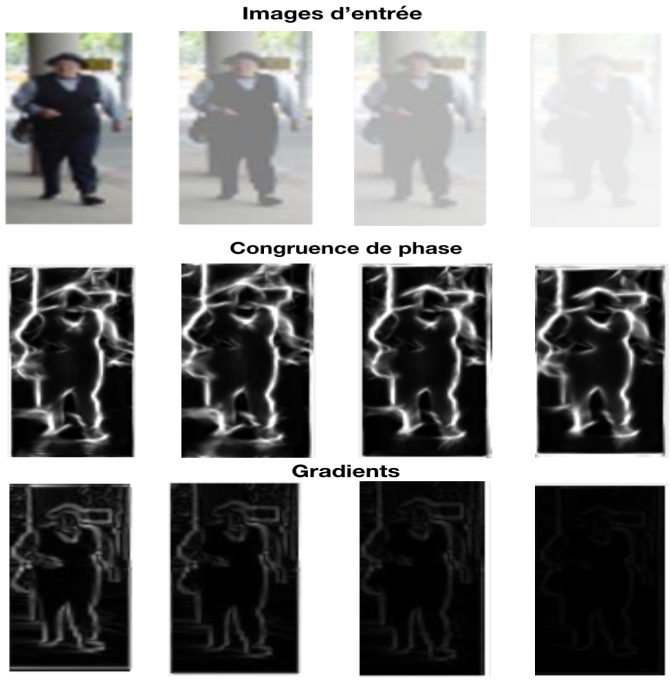


FIGURE 1.10 – Comparaison de la congruence de phase avec le gradient

Etant donnée une image $I(x)$, l'expression de la congruence de phase est donnée par :

$$PC(x) = \frac{E(x)}{\epsilon + \sum_n A_n} \tag{1.2}$$

où $E(x)$ est l'énergie locale, A_n l'ensemble des amplitudes des composantes de Fourier et ϵ une petite quantité pour éviter la division par zero. L'expression de l'énergie locale est donnée par [21] :

$$E(x) = \sqrt{F(x)^2 + F(x)_H^2} \tag{1.3}$$

avec $F(x)$ le signal filtré de $I(x)$ et $F(x)_H$ la transformée de Hilbert de $F(x)$ (soit un décallage de 90° de $F(x)$). L'orientation de la phase est donnée par :

$$\phi(x) = tan^{-1}\left(\frac{F_H(x)}{F(x)}\right) \tag{1.4}$$

Dans la pratique, la congruence de phase peut être estimée par convolution de l'image par une paire quadratique des filtres de log-Gabor.

Ragb et al. [21] ont proposé le descripteur CHOP en se basant à la fois sur les avantages de la congruence de phase mais aussi sur la méthode d'extraction utilisée par le descripteur HOG pour faire de la détection de personnes. Les auteurs ont présenté des résultats qui montrent la supériorité de ce dernier sur le descripteur HOG dans certaines situations bien précises. La démarche est presque identique à celle de HOG à l'exception de l'utilisation des orientations de la congruence de phase à la place du gradient. La figure 1.11 résume les différentes étapes de l'extraction des vecteurs CHOP.

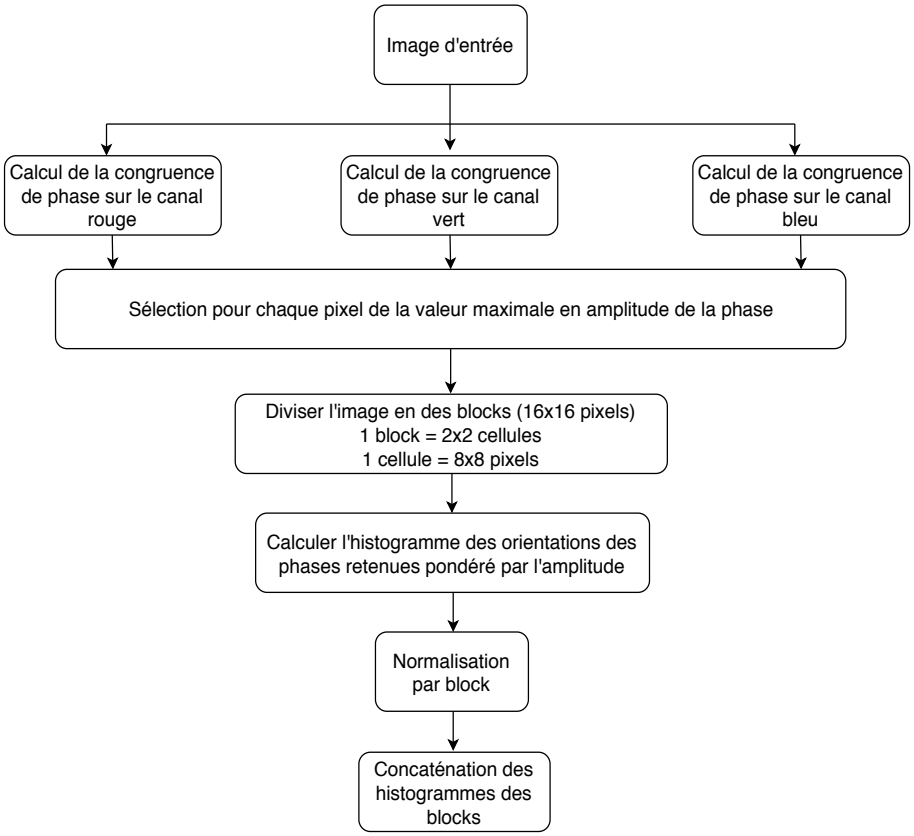


FIGURE 1.11 – Différentes étapes de l'extraction de CHOP [21]

1.3.4 Les filtres de convolution

L'extraction de caractéristiques à base des filtres de convolution est, on peut le dire, ce qui se fait de mieux actuellement dans l'état de l'art. Les filtres de convolution sont utilisés dans le domaine des images pour mettre en évidence un certain nombre d'informations non perceptibles dans une image brute. La convolution d'une image par un filtre avec un noyau donné, est une opération matricielle, qui à partir d'une position donnée dans l'image, calcule la somme du produit élément-à-élément du noyau avec les voisins du pixel central (eq 1.5).

$$z(n_1, n_2) = \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} x(k_1, k_2) y(n_1 - k_1, n_2 - k_2) \tag{1.5}$$

Le résultat de l'opération est la réponse de cette région de l'image par rapport au filtre. De nombreux filtres existent et sont utilisés dans plusieurs descripteurs pour l'extraction de caractéristiques primaires, tel que le filtre de Sobel qui est utilisé pour la mise en évidence des contours dans une image. A la différence des descripteurs traditionnels, comme ceux présentés précédemment, les méthodes basées sur les filtres de convolution, produisent des cartes de caractéristiques ("feature maps" en anglais) qui sont par la suite exploitées de manière spéciale pour les détections. Les filtres de convolution possèdent plusieurs avantages qui leur confèrent une certaine efficacité par rapport aux autres méthodes traditionnelles : ils permettent de conserver les caractéristiques spatiales entre les pixels de l'image puisque la carte de caractéristiques produite est également une image dont chaque pixel représente la caractéristique dans la même position

au niveau de l'image précédente. Dans les architectures de réseaux de neurones convolutifs, où ils sont majoritairement utilisés dans le but d'extraire des caractéristiques, les valeurs des filtres peuvent être apprises automatiquement par le même procédé d'apprentissage utilisé pour ajuster les poids des réseaux de neurones tels que la rétropropagation du gradient. En outre, les filtres de convolution peuvent être disposés en couche de manière hiérarchique ; cette caractéristique leur permet d'être très efficace dans la détection de motifs plus complexes dans les images et d'obtenir des représentions plus abstraites des images.

De plus en plus, les architectures de réseaux de neurones convolutifs sont utilisées pour des tâches de détection et de classification d'images. Ainsi, l'architecture CNN (Convolutional Neural Network) est un réseau constitué de plusieurs couches convolutives entemêlées de couches de pooling. Chaque étage de l'architecture est constitué d'une convolution suivie d'un pooling. L'entrée de la première couche est utilisée pour les données brutes dont les cartes de caractéristiques issues à la sortie servent d'entrée pour la couche suivante. Ainsi de suite jusqu'à la dernière couche qui sert à faire une transformation non linéaire des caractéristiques issues de l'avant-dernière couche. Cette couche peut servir de classifieur dans une tâche de classifaction d'image. La figure 1.12 présente un exemple de réseau CNN à trois couches.

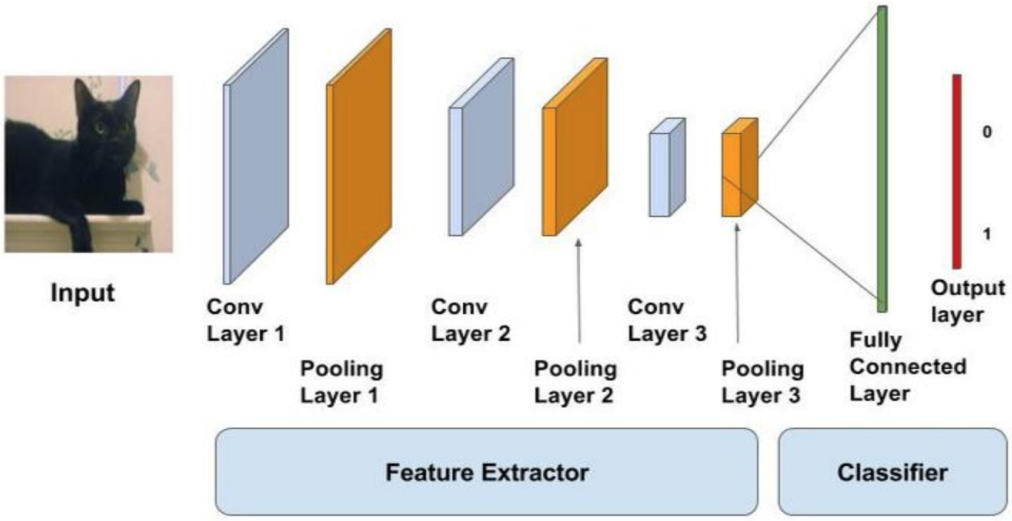


FIGURE 1.12 – Exemple de réseau CNN. soucre : `https:/www.learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras`

Dans un réseau CNN, plusieurs filtres sont généralement utilisés au niveau de chaque couche et la fonction maxPooling est utilisée au niveau des couches de Pooling pour réduire la taille spatiale de la carte de caractéristiques et éviter le surapprentissage au niveau du réseau. Des versions améliorées du réseau CNN de base ont été utilisées pour améliorer les performances et surtout le temps de calcul. On peut ainsi citer les travaux de Girshick et al. [22, 23, 24] sur des modèles R-CNN, Fast R-CNN et Faster R-CNN pour la détection et la segmentation de scène.

1.4 Les algorithmes d'apprentissage

Il existe en "machine learning" plusieurs algorithmes qui permettent d'apprendre des modèles à partir d'un ensemble de vecteurs de caractéristiques d'un jeu d'apprentissage. Pour les tâches de détection d'objets, bon nombre de ces algorithmes ont été associés aux descripteurs. Nous présenterons dans cette sous-section trois de ces algorithmes de classification à savoir : les machines à vecteurs de support (SVM), les classifieurs bayésiens et l'algorithme Adaboost.

1.4.1 L'algorithme Adaboost

Adaboost [25, 26] est un puissant algorithme d'apprentissage automatique. Il combine un nombre T de classifieurs faibles h_t en un classifieur fort H. Les classifieurs sont choisis pour minimiser l'erreur sur les échantillons de l'ensemble d'apprentissage. L'erreur de classification ϵ_t sert à évaluer la confiance α_t dans la décision de classification. A la sortie de chaque itération, les données d'apprentissage mal classées obtiennent un poids élevé pour permettre à l'itération suivante de forcer le classifieur faible à mieux les apprendre. Pour la phase de classification d'une donnée inconnue, chaque classifieur faible h_t donne une valeur de confiance α_t dont la somme pour tous les classifieurs donne la valeur de décision finale de classification. Plusieurs versions de l'algorithme ont été proposées dans la littérature pour de multiples tâches de classification.

1.4.2 Les classifieurs bayésiens

La résolution des problèmes de classification repose le plus souvent sur l'émission d'hypothèses qui ramènent le problème dans un cadre bayésien. Un classifieur bayésien ou estimateur bayésien est un type de classifieur probabiliste basé sur le théorème de Bayes. Il regroupe un ensemble d'algorithmes qui se fondent tous sur le même théorème avec des variantes. Le but d'un classifieur bayésien est d'apprendre la distribution des valeurs des paramètres du modèle. Pour ce faire, une distribution a priori (représentation des hypothèses émises) des valeurs des paramètres est fournie à l'algorithme qui effectue l'apprentissage à travers la règle de Bayes (voir equation 1.6). L'apprentissage de la distribution se fait à posteriori à travers une estimation de la vraissemblance. L'estimation de cette vraissemblance est une tâche difficile pour laquelle plusieurs algorithmes existent. On peut citer les algorithmes EM (Expectation-Maximization) de Dempster et al. [27], les méthodes de fenêtre de Parzen [28] et les méthodes de Monte Carlo par chaînes de Markov. Ces classifieurs présentent l'avantage d'éviter le surapprentissage mais nécessitent cependant des hypothèses fortes sur les distributions des données, ce qui les rend difficilement applicables aux problèmes réels.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{1.6}$$

où $P(A|B)$ désigne la probabilité conditionnelle de A sachant B.

1.4.3 Séparateurs à Vastes Marges

Le SVM (Les Séparateurs à Vastes Marges ou Machine à vecteurs de support) [29] est l'un des algorithmes d'apprentissage automatique supervisés le plus utilisé pour à la fois des problèmes

de classification et de regression. Le but du SVM est de trouver ou de déterminer un hyperplan qui permettra de mieux séparer un jeu de données en deux ou plusieurs classes (voir figure 1.13). Les vecteurs de support, d'où est tiré le nom de l'algorithme, sont l'ensemble des points les plus proches de l'hyperplan. L'algorithme cherche alors à optimiser la distance entre les échantillons les plus proches (les vecteurs supports) et l'hyperplan. Dans le cas d'une classification binaire, comme la détection de personnes, le jeu de données est composé de données positives représentant les échantillons à détecter et des données négatives qui peuvent tout contenir sauf les échantillons à détecter. La classification de données de test consiste à calculer la distance entre le vecteur de test et l'hyperplan. Cette distance donne l'information sur le côté où se trouve la donnée de test et aussi une mesure de la fiabilité de la décision prise. En effet, plus la distance est petite moins la décision est tranchée car la donnée est très proche des deux classes.

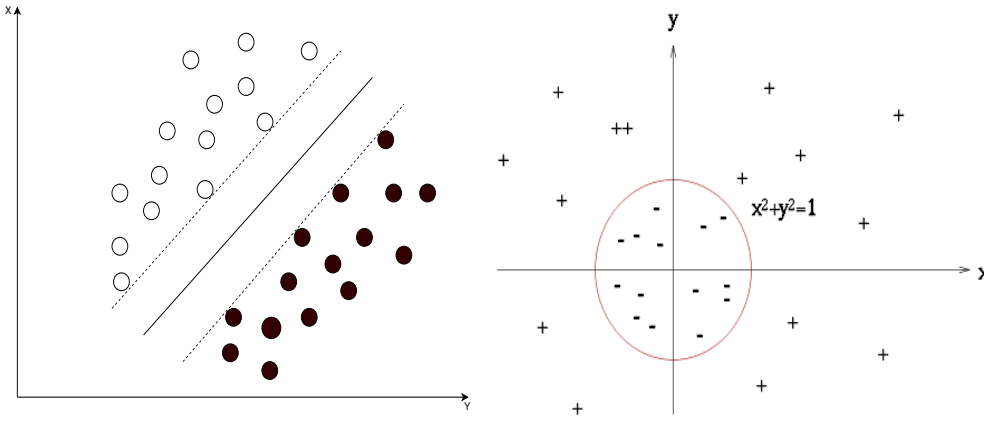


FIGURE 1.13 – Illustration du SVM pour une séparation linéaire simple à gauche et non linéaire à droite.

Pour des données linéairement inséparables, la recherche de l'hyperplan se fait dans un espace de plus grande dimension en appliquant une transformation non-linéaire ϕ sur ces denrières. Ainsi, l'équation de l'hyperplan peut s'écrire sous la forme :

$$h(x) = w^T \phi(x) + w_0 \tag{1.7}$$

avec w le vecteur de poids qui vérifie la condition $l_k h(x) > 0$ et l_k les labels.
Cette expression montre bien qu'il s'agira d'un produit scalaire entre deux vecteurs dans l'espace de grande dimension, ce qui peut être couteux en temps de calcul. Une solution pour contourner ce problème est l'utilisation d'une astuce appelée "Kernel trick" qui consiste à remplacer le produit scalaire par une fonction noyau qui vérifie $K(x_i, x_j) = \phi(x_i)^T.\phi(x_j)$. $K(x_i, x_j)$ doit respecter le théorème de Mercer, c'est à dire qu'elle doit être symétrique et semi-définie positive. L'équation 1.7 devient :

$$h(x) = \sum_{k=1}^{p} \alpha_k^* l_k K(x_k, x) + w_0 \tag{1.8}$$

Quelques fonctions noyau utilisées fréquemment sont :
- le noyau linéaire : $K(x_i, x_j) = x_i^T.x_j$;
- le noyau polynomial : $K(x_i, x_j) = (x_i^T.x_j + 1)^d$;
- le noyau gaussien : $K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$;

1.5 Les méthodes de détection

Le modèle classique de détection se présente sous l'architecture descripteur-classifieur. Les différentes méthodes proposées dans la littérature différent l'une de l'autre à la fois par le ou les descripteur(s) utilisé(s) mais aussi par le classifieur employé pour la construction du modèle. Dans notre contexte, le rôle du classifieur est de prédire si un vecteur de caractéristiques représente une image de personne. Pour des images d'un environnement donné, contenant des personnes et d'autres objets, des stratégies de recherche sont nécessaires pour détecter et localiser les entités recherchées. Nous avons étudié quelques types de stratégies dont nous donnons un bref aperçu dans cette section.

1.5.1 Méthode de fenêtres glissantes : sliding windows

La fenêtre glissante est la méthode de base utilisée pour la détection et la localisation d'objets dans une scène. L'idée est simple et consiste à parcourir entièrement une image avec une fenêtre de taille donnée. Les images des régions balayées par la fenêtre glissante sont soumises à une classification binaire. Les fenêtres sont extraites à différentes échelles afin d'inclure des personnes de différentes tailles. Cette méthode naïve est la plus sûre car assurant un parcours complet de l'image, mais présente naturellement un désavantage lié au temps de calcul avec une complexité $O(n^4)$ pour une image de taille $n \times n$. Dans une image, il y a généralement moins de personnes, d'objets que d'éléments de l'arrière plan. Pour réduire le nombre de fenêtres à examiner et ainsi améliorer le temps de calcul, Lin et al. [30], Lu et al. [31] précèdent la détection des fenêtres candidates d'une étape de soustraction d'arrière plan afin d'isoler les éléments du premier plan. Les fenêtres sont alors extraites uniquement dans les régions isolées. Cette approche nécessite néanmoins l'utilisation d'un algorithme robuste de soustraction d'arrière-plan. A la fin de la classification, une personne peut être détectée dans plusieurs fenêtres. La méthode NMS (Non Maximum Suppression) est généralement utilisée pour fusionner les fenêtres.

1.5.2 La méthode "Efficient sub-window search"

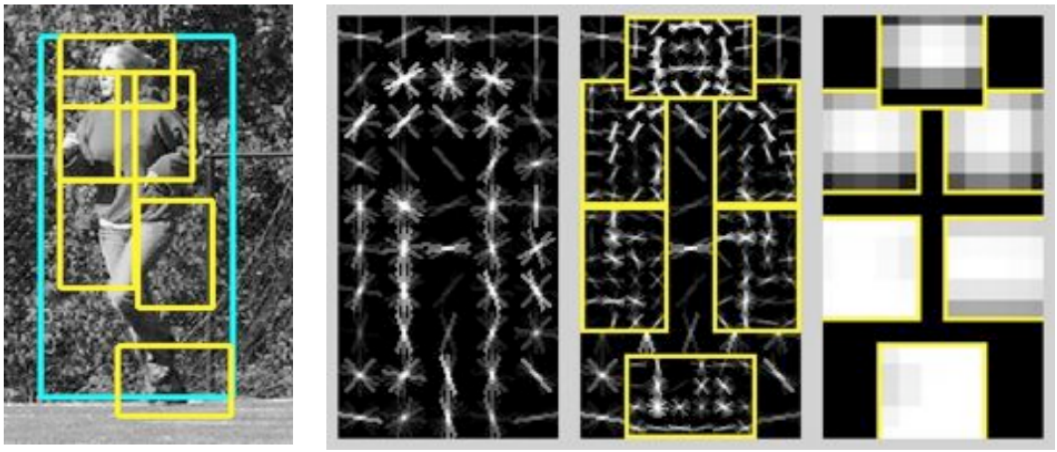
La méthode ESS est une méthode de recherche de fenêtres contenant un objet à détecter sans faire une recherche exhaustive dans l'image. Elle utilise la méthode "branch and bound" pour explorer uniquement les fenêtres avec des scores élevés. En partant de l'hypothèse que les fenêtres ont une forme fixe (rectangulaire ou carré, par exemple), la méthode ESS définie une stratégie efficace de réduction du domaine de recherche des objets en rejetant au fur et à mesure les régions de l'image ayant un faible score, donc une faible probabilité de contenir l'objet recherché. La méthode consiste à diviser à chaque itération les rectangles ayant un grand score de classification en deux rectangles, qui seront à l'itération suivante évalués. Plusieurs versions de la méthode existent et diffèrent l'une de l'autre dans la stratégie de recherche, au niveau des variables utilisées, au niveau de la métrique de calcul du score de similarité, etc. On peut citer $\chi^2 - ESS$ [32] qui utilise la distance χ^2 dans la phase de calcul du score des fenêtres.

1.5.3 Les méthodes "Region Proposal"

Il s'agit de méthodes ayant pour but de déterminer des régions susceptibles de contenir un objet recherché. Elles identifient les objets potentiels dans l'image à l'aide de la segmentation. Ainsi, contrairement à la méthode de fenêtres glissantes, les méthodes de proposition de régions vont générer des régions segmentées, dans lesquelles les recherches pourront avoir lieu pour la détection. Cela réduit énormément le nombre de fenêtres à examiner. Ells sont régulièrement utilisées dans les architectures de réseaux convolutifs pour déterminer les régions où les filtres doivent être appliqués. Une de ces méthodes est le "Selective Search", développée par Uijlings et al. [33] et qui a été utilisée dans les réseaux comme R-CNN et Fast-RCNN. Elle est basée sur le regroupement hiérarchique de régions similaires en fonction de la couleur, de la texture, de la taille et de la compatibilité des formes. La recherche commence par une sur-segmentation de l'image en fonction de l'intensité des pixels en utilisant la méthode de segmentation basée sur un graphe de Felzenszwalb et Huttenlocher [34]. La sortie de cette étape est une image grossièrement segmentée dans laquelle un objet peut être segmenté en deux régions. En se basant sur cette première segmentation, la méthode "selective Search" définit une liste de toutes les régions identifiées puis dans une boucle, essaie de regrouper les régions adjacentes en fonction de leur similarité calculée à partir des caractéristiques de couleurs, de textures et de tailles. A la fin des itérations, un nombre réduit de grandes régions est obtenu et peuvent servir d'entrée à l'algorithme de classification.

1.5.4 Le modèle "Deformable Part Models"

L'hypothèse derrière la méthode DPM (Deformable Part Models) de Felzenszwalb et al. [35] est qu'un modèle unique de personne ne peut être assez efficace pour la détection surtout en présence d'occultations. Pour faire une détection effective des personnes même partiellement visibles dans l'image, ils proposent de construire en plus du modèle de la forme complet, des modèles de différentes parties qui sont spatialement liées à la fenêtre du modèle de la forme complète (par exemple un modèle de la tête, des pieds, etc.). A chaque modèle de partie est lié un filtre spécifique qui servira à calculer le score de détection dans l'image. L'évaluation d'une fenêtre repose donc sur les scores du filtre de la forme complète ainsi que ceux des filtres des autres parties. Dans l'évaluation de leur approche, les auteurs ont utilisé HOG comme descripteur et une forme spécifique de SVM qu'ils ont nommé LSVM (Latent SVM) pour apprendre les différents modèles de partie. Ce modèle de détection est utilisé conjointement avec les méthodes de recherche présentées plus haut, pour l'identification des fenêtres à évaluer. La figure 1.14 montre un exemple de détection de personne avec la visualisation des gradients et des filtres utilisés pour les différentes parties.



a b c d

FIGURE 1.14 – Exemple de détection de personne avec le DPM [35] :
a : Détection d'une personne
b : Filtre entier
c : Filtres des parties
d : Modèles des parties

1.5.5 Les approches par BOVW psycho-visuels

Cette approche propose d'intégrer les modèles de perception humaine dans le processus d'interprètation des scènes tel que pour les tâches de reconnaissance de personnes, d'objet, etc. Il s'agit de partir de la carte de saillance visuelle pour guider le choix des zones de la scène à interprèter. Le principal objectif est de réduire le temps de calcul nécessaire pour la recherche en comparaison des approches basées sur les fenêtres glissantes.

Gonzalez et al. dans [36] ont proposé, dans cette catégorie, une approche qui intégre dans le paradigme de "sac-de-mots visuels" l'utilisation des cartes de saillances visuelles. Plus précisément, ils proposent, à l'étape de construction de l'histogramme des occurences de mots-visuels, de pondérer le vote de chaque vecteur de caractéristiques par le score de saillance visuelle qui correspond à l'endroit de la scène où le point d'intérêt qu'il décrit a été détecté. Une normalisation L1 peut être employée sur le vecteur d'occurence obtenue. Dans le même papier, ils propsent une méthode de génération automatique de dictionnaire et de détermination du nombre optimal de mots visuels nécessaire pour la description de la scène à chaque résolution spatiale de la scène associée à un niveau dans la pyramide des images. Cette méthode se base également sur les informations de saillance visuelle de la scène. Gonzalez et al. ont montré que cette méthode surpasse l'approche DPM présentée plus haut et donne des résultats satisfaisants sans un grand impact en terme de temps de calcul.

1.5.6 Synthèse des méthodes de la littérature

Un nombre important de travaux existe dans la littérature sur la détection de personnes. Nous résumons dans le tableau 1.1 quelques uns de ces travaux sans entrer dans les détails des descripteurs et classifieurs utilisés. Il ne s'agit pas d'une liste exhaustive des méthodes existantes. A travers ce tableau, nous montrons qu'à un moment donné, la communauté scientifique s'est

focalisée sur le descritpeur HOG et l'algorithme SVM pour cette tâche de détection de personne.

Tableau 1.1 – Synthèse des approches de détection de personnes dans une image

Références	**Descripteur**	**Classifieur**	**Observations**
Shayhan et al. [37]	HOG	SVM Linéaire	Etape initiale d'extraction de régions et gestion d'occultations
Ragb et al. [21]	CHOP-LUV	SVM Linéaire	Description des Orientations des Phases. Robuste au changement de luminosité
Xiao et al. [38]	HOG	LDA avec L1-norm	Optimiser le temps d'exécution en utilisant le LDA (Linear Discriminant Analysis)
Dalal et al. [17]	HOG	SVM Linéaire	Méthode de base de la détection avec HOG
F. Suard et al. [39]	HOG	SVM	Détection dans images infrarouges + stéréovision
Benenson et al. [40]	chnFtrs et DPM	discret Adaboost	Accélération de la détection par approximation des caractéristiques pour différentes échelles
Chu et al. [41]	mspw-DPM et HOG-based BRIEF	-	Particule adaptative pour générer des fenêtres potentiellement positives pour accélérer la détection
Sheng et al. [42]	HOG-LBP	Gentle AdaBoost	Combinaison de l'information de contours et de texture
Wang et al. [43]	HOG-LBP	SVM	Gestion des occultations en faisant de la segmentation à partir de la carte de probabilité des réponses du classifieur
Yunsheng et al. [44]	HOG-color-barShape (HOG-III)	Grammar Model et Poselet Model	fusion de caractéristiques
Girshich et al. [22]	R-CNN	-	CNN combiné avec une méthode de proposition de régions pour accéler le temps de calcul
Redmon et al. [45]	YOLO	-	Traite le problème de détection comme un problème de régression

1.6 Résultats expérimentaux

Dans cette sous-section, nous allons présenter quelques résultats d'évaluation de certains algorithmes sur des données publiques de détection de personnes. Dans un premier temps, nous présentons des résultats de tests de combinaisons de descripteurs existants de la littérature afin de les comparer aux résultats directement issus de la littérature. Dans un second temps, il sera présenté une comparaison dans le contexte de détection de personne dans la nuit.

1.6.1 Bases de données d'évaluation

- La base INRIA

La base "INRIA", disponible à l'adresse `http://lear.inrialpes.fr/data`, contient 1239 images en couleur de personnes augmentées de 1239 images issues de la transformation symétrique suivant l'axe vertical de ces dernières. La base contient également des données négatives au nombre de 1218 qui ne contiennent aucun humain mais une grande diversité de paysages et d'environnements urbains et ruraux et certains objets comme des bicyclettes ou des voitures. Pour l'évaluation des méthodes, 1126 images positives et 453 images négatives sont fournies. Toutes les images positives ont été redimensionnées à la taille 64×128 et contiennent des personnes en position debout avec différentes orientations. Les personnes portent des vêtements de différentes couleurs et textures. Les photos sont réalisées dans différentes conditions météorologiques (hiver, été). Les images négatives sont de grandes dimensions, ce qui permet de sélectionner au hasard dans ces dernières plusieurs images de taille 64×128. Ainsi, pour l'apprentissage, dans une grande image négative, on peut sélection 10 images à la taille 64×128 et ainsi augmenter à 12180 le nombre de données négatives dans la base.

- La base CVC-14

Cette base d'images, disponible à l'adresse `http://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset/`, est composée de deux ensembles de séquences. Ces séquences sont nommées d'après les ensembles de jour et de nuit, qui font référence au moment de la journée où l'acquisition des données est faite. Le jeu d'apprentissage contient 3695 images de jour et 3390 images de nuit, avec environ 1500 images labélisées. Le jeu de test contient environ 700 images pour les deux séquences avec environ 2000 personnes pour les images de jour et 1500 personnes pour celles de la nuit. Cette base servira à comparer la robustesse des descripteurs face au changement de luminosité.

1.6.2 Détection dans un contexte nocturne

Ragb et al. dans leur travail sur la détection de personnes [21] ont montré la supériorité d'un certain nombre de descripteurs basés sur la congruence de phase dont le descripteur CHOP sur les descripteurs HOG, PHOG et uLBP. Il est à souligner que dans leur travail, les résultats ont été obtenus à partir de modèles appris sur des données de la base INRIA et testés sur la même base. Autrement dit, la performance de leur descripteur sur les autres a été prouvée dans des conditions

de luminosité du jour. Dans cette section, nous souhaitons étudier la robustesse des descripteurs HOG et CHOP pour des images de nuit. En effet, comme mentionné dans les sections précédentes, le contexte des travaux de cette thèse nous amène à rechercher des descripteurs robutes pouvant être utilisés dans différentes conditions lumineuses de détection. Nous cherchons donc à créer un modèle unique à partir d'images prises dans des conditions de jour ou de nuit et à tester la détection dans les différentes conditions. Pour ce faire, nous avons effectué les tests suivant deux scénarii à savoir :

- **Scénario 1 :** Apprentissage d'un modèle avec des images de jour et des tests avec des images de nuit ;

- **Scénario 2 :** Apprentissage d'un modèle avec des images de nuit et des tests dans les mêmes conditions.

Un troisième scénario peut être envisagé mais a été déjà pris en compte dans les travaux de Rage et al. [21]. Il s'agit de l'apprentissage et du test sur des images de jour. Les modèles de personnes sont construits avec le type nu-SVC de l'algorithme SVM. La valeur du paramètre nu du SVC est déterminée durant la phase d'apprentissage. La courbe rappel-précision est le critère d'évaluation des performances que nous utilisons. Avec ce critère, la méthode dont la courbe se trouve plus en haut présente la meilleure performance.

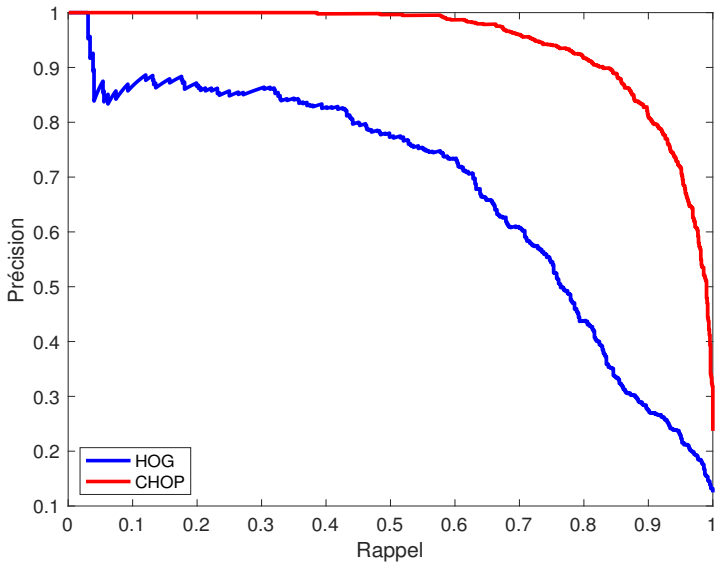


FIGURE 1.15 – Courbe pour le test selon scénario 1 : Apprentissage de jour et test de nuit

La figure 1.15 montre le résultat pour le premier scénario. Pour ce dernier, nous avons créer un modèle de personne par descripteur à partir de la base INRIA et les tests ont été effectué à partir des images de la base CVC-14. On note que le descripteur CHOP est plus robuste que le descripteur HOG. Il est donc bien adapté pour une détection dans des images nocturnes en infra-rouge à partir de modèle appris sur des images de jour. Le descripteur HOG ne donne pas une bonne performance dans ces conditions.

Le test du scénario 2 visent à déterminer les performances des descripteurs si les modèles

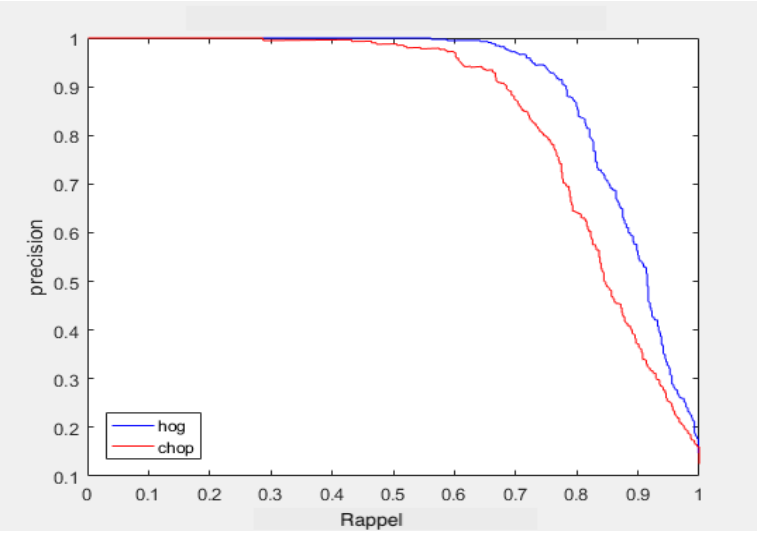


FIGURE 1.16 – Courbe pour le test selon scénario 2 : Apprentissage de nuit et test de nuit

étaient appris et testés avec des images de nuit. La figure 1.16 présente les courbes les résultats à cet effet et obtenus en utilisant la base CVC-14. De ces résultats nous pouvons déduire que le descripteur HOG est meilleur que le descripteur CHOP.

1.7 Conclusion

Ce premier chapitre est consacré à l'étude de différents descripteurs de formes utilisés dans la littérature pour diverses tâches de détection. Le travail présenté dans ce chapitre comporte deux parties. La première partie traite des descripteurs utilisés le plus souvent pour la détection de personnes, d'objets. Le but premier de notre travail est d'étudier la robustesse des descripteurs les plus utilisés face au changement de condition d'éclairage. Dans cette catégorie, nous avons brièvement présenté les descripteurs locaux tels que le descripteur Haar-like et les descripteurs de points d'intérêts SIFT et SURF. Mais ce sont les descripteurs globaux, qui sont les plus utilisés pour la tâche de détection de personne, qui ont attiré notre attention. Il s'agit des descripteurs HOG, LBP et CHOP. En nous basant sur les travaux récents de Ragb et al., nous avons testé et comparé les descripteurs HOG et CHOP suivant deux scénarii. Nous avons utilisé les bases de données publiques INRIA et CVC-14 pour réaliser ces deux scénarri. En apprenant un modèle à partir des images prises dans les conditions d'éclairage du jour et en testant uniquement sur des images de nuit, nous avons pu observer que le descripteur CHOP donne de meilleure performance comparé au descripteur HOG. Autrement dit, pour avoir de bonnes détections dans des conditions de luminosité difficiles avec un modèle unique entrainé sur des images de jour, le descripteur CHOP est le mieux approprié. A l'opposé, en apprenant un modèle avec les images de nuit et en testant dans les mêmes conditions, le descripteur HOG prend le dessus. Notre étude nous amène à conclure que le descripteur CHOP résiste mieux à la variation de la luminosité mais qu'à conditions d'apprentissage et d'entraînement égales, le descripteur HOG le mieux adapté.

Ce chapitre a ainsi :

- présenté une description et comparison de descripteurs populaires et performants de la littérature pour la tâche de détection de personnes dans des conditions de luminosité variantes ;

- permis la mise en place d'un système de détection des entités tels que les piétons. Ce système a servi dans le cadre applicatif du projet.

Chapitre 2

Détection des événements rares : L'état de l'art

> Le contraire d'une vérité banale, c'est une erreur stupide. Le contraire d'une vérité profonde, c'est une autre vérité profonde.
>
> Niels Bohr

Sommaire

2.1 Introduction

La sécurité des lieux publics tels que : les routes, les aéroports, les centres commerciaux, les stations de métro, etc, constitue de nos jours, un énorme défi. Avec l'explosion de l'utilisation des caméras de surveillance au quotidien, le besoin d'offrir aux agents de sécurité un support de décision rapide et efficace en cas de problème, est plus que jamais exprimé. Un tel besoin, peut s'expliquer par la limitation de la capacité humaine à assurer une surveillance constante et rigoureuse de plusieurs lieux simultanément [46]. Pour relever ce défi, les communautés de vision par ordinateur et d'intelligence artificielle ainsi que d'autres avant eux, ont abordé le problème, en proposant des solutions, il y a quelques dizaines d'années [47, 48, 49]. Le but était de mettre au point des algorithmes fiables pouvant identifier le plus rapidement possible des situations suspectes afin d'émettre des alertes.

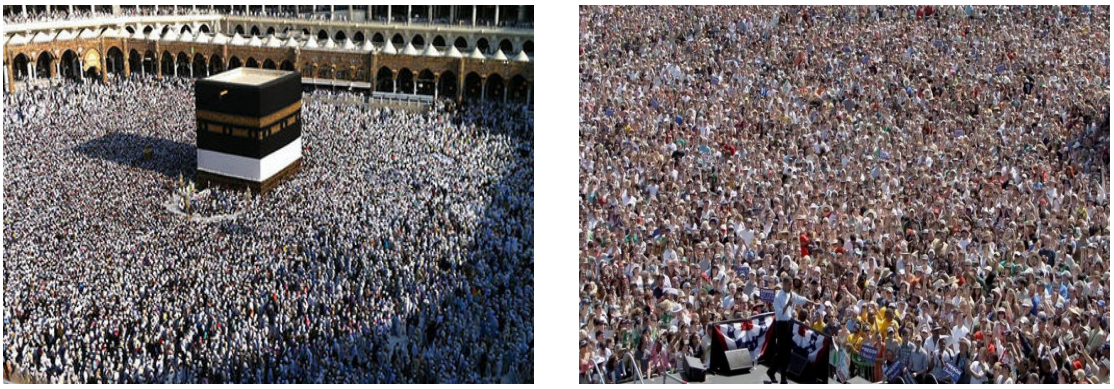
L'une des manières d'aborder ce problème est de considérer les situations suspectes, comme étant des faits et gestes non fréquents dans un environnement donné. On peut prendre les exemples de mouvements d'évacuation d'un lieu, d'abandon de colis, de piétons sur une chaussée, d'accidents de circulation, de déclenchement de feu, etc. Ce genre de comportement n'est pas fréquemment observable, mais apparaît juste dans le temps. Dans cette thèse, nous nous positionnons dans ce régistre qui revient à la détection de "outliers" par rapport à un modèle d'événement appris. Ainsi, la notion d'événements "rares" sera plus appropriée que celle d'événements anormaux, abondamment utilisée dans la littérature. Il s'agit d'un problème ouvert car la notion d'anormalité et son interprétation demeure très subjective. Ainsi, en fonction des applications visées, la nature des événements varie fortement ce qui rend difficile une comparaison objective des approches proposées dans l'état de l'art.

Une grande partie des travaux réalisés sur le sujet s'intéresse plus généralement à l'analyse de scènes et plus spécifiquement à l'analyse de foules. Plusieurs thématiques sont regroupées sous le thème "analyse de scènes" : de l'analyse du comportement des foules en présence de dangers à la détection d'anormalités dans une scène en passant par la segmentation de foule. Plusieurs domaines scientifiques tels que la physique, l'infographie ("computer graphic") et la vision par ordinateur, se sont intéressés à la résolution de ce problème. Par scène de foule, il faut comprendre un environnement où se déplacent des personnes avec des mouvements structurés ou non. Une scène est souvent catégorisée en fonction de la densité de foule qui s'y trouve (faible, moyenne et forte) mais aussi en fonction des mouvements qui s'y déroulent. Les différentes approches se différencient par rapport au type de scène qu'elles traitent et des objectifs qu'elles poursuivent. On peut trouver deux grandes catégories d'approches : celles qui traitent des scènes de foule à moyenne et forte densités et celles qui s'attachent aux scènes à faible densité. Dans une scène à moyenne ou forte densité de foule, l'objectif est de faire une analyse globale de la scène pour détecter le début et la fin d'un mouvement inhabituel. En complément, pour les scènes à faible densité, non seulement l'objectif est de détecter le début et la fin des événements inhabituels, mais également de les localiser précisément dans la scène avant d'en étudier la nature éventuellement.

Dans ce chapitre, nous ferons la revue des grandes approches de la littérature qui ont abordé

les deux catégories de scènes. Ces approches se sont, pour certaines, inspirées de la physique plus précisément de la mécanique des fluides, et d'autres sont purement basées sur les techniques propres à la "vision par ordinateur". Une dernière catégorie que nous aborderons, concerne les approches basées sur les réseaux de neurones convolutifs, qui ne sont qu'à leur début dans le domaine de machine learning et pour la thématique que nous traitons.

2.2 Détection d'événements rares dans une scène à moyenne et forte densité de foule



a b

FIGURE 2.1 – Exemple de scène de foule.
a : foule autour de la Ka'ba à la Mecque dont le mouvement de rotation est structuré
b : foule lors d'un meeting politique dont le mouvement est non structuré

Une scène de foule est définie par rapport à la densité des entités (personnes, voitures, motocyclistes, etc) qui s'y trouvent. Les scènes de foules présentent une complexité de modélisation qui augmente en fonction de la structuration des mouvements des entités. On peut distinguer des foules dont les mouvements sont bien structurés et des foules aux mouvements non structurés (voir Fig 2.1). On parle de scènes de foules dont les mouvements sont structurés quand les entités présentes suivent des mouvements presque uniformes (exemple du tour de la Ka'ba à la Mecque). A contrario, les scènes de mouvements non structurés, présentent des mouvements désordonnés pour lesquels il est difficile d'identifier des schémas de mouvements collectifs. Cette section est dédiée à la présentation des approches qui visent la détection d'événements rares dans des foules à moyenne et forte densité. Pour de telles scènes, les approches ne visent pas à modéliser le comportement des entités mais plutôt à réaliser une analyse globale. Le but des travaux initiaux n'était pas une détection d'événements rares, mais plutôt une modélisation des foules à des fins de simulation. Que ce soit en physique ou en infographie ("computer graphic"), la simulation vise une représentation réaliste des foules, de leur comportement soit à des fins d'études de plans d'évacuation, soit pour des jeux vidéos, ou autres. Plus tard, en "vision par ordinateur", la thématique de détection d'événements rares ou anormaux (selon les considérations des auteurs) a concerné la détection du début et de la fin d'un comportement déviant dans une scène de foules. Ces méthodes s'inspirent aussi bien de la physique que des techniques propres au domaine de la

vision.

2.2.1 Modélisation de la dynamique de foule en physique

L'étude des mouvements de foules a débuté en physique il y a de cela plusieurs années. Au départ, ces différentes études étaient menées dans le but de décrire la formation et le mouvement des foules mais aussi l'interaction entre les entités présentes dans la foule. Comme mentionné ci-dessus, la complexité de la modélisation de la foule augmente avec la complexité structurelle de la foule mais aussi avec sa densité. Dans la littérature, deux niveaux d'études des foules se rapportant à la densité se dégagent : le niveau macroscopique et le niveau microscopique.

Au niveau macroscopique, il est presque impossible de se focaliser sur chaque entité dans la scène afin d'en étudier le comportement. Ainsi, les approches proposées en physique pour faire la modélisation à ce niveau, traitent la foule comme un fluide avec des particules dont l'écoulement est étudié à l'aide de méthodes issues de domaines variés tels que la mécanique des fluides, la thermodynamique, la statistique, etc. Parmi les travaux ayant aborder le problème pour cette catégorie de foule, on peut citer le modèle "continuum-based" [50] de Hughes qui représente les piétons comme un champ de densité continue décrivant la dynamique de la foule. Il propose des équations différentielles basées sur trois hypothèses qu'il a émis pour la modélisation de la dynamique de la foule. Le modèle proposé n'est valable que pour des foules composées de personnes ayant un but ou une destination. On peut prendre l'exemple de l'évacuation d'un bâtiment où le but de chacun est d'aller à l'extérieur du bâtiment en passant par les issues de secours. Plus tard, Treuille et al. [51] ont proposé un modèle en temps réel, basé sur le modèle "continuum" de Huges. Leur modèle intègre simultanément la navigation globale avec des obstacles mobiles tels que d'autres personnes, résolvant ainsi efficacement le mouvement de grandes foules sans avoir besoin de se focaliser sur l'évitement de collision.

Pour parvenir au modèle, ils ont formulé quatre hypothèses, basées sur celles de Hughes, que nous résumons ci-dessous :

- toute entité dans la foule a un but ou une destination à atteindre. Soit $G \subset \mathbb{R}^2$ l'ensemble des buts des entités de la foule. Elles éprouvent des difficultés à marcher à contre-courant, ce qui est proportionnel à la densité locale de la foule, (Hypothèse 1)

- les personnes se déplacent toujours avec leur vitesse maximale possible. Soit v cette vitesse, x la position de l'individu à un instant donné et θ la direction qu'il prend. La vitesse est une fonction des deux variables telles que : $v = f(x, \theta)n_\theta$ avec $n_\theta = [cos\theta, sin\theta]^T$ un vecteur unitaire qui pointe dans la direction choisie, (Hypothèse 2)

- il existe une fonction d'inconfort g telle que toute personne a une préférence d'être à un endroit x plutôt qu'à un autre endroit x' tel que $g(x') > g(x)$. Cette fonction a été utilisée pour faire l'évitement d'obstacles, (Hypothèse 3)

- il existe plusieurs chemins possibles que peuvent emprunter les entités constituant la foule. Soit $\prod$ cet ensemble. L'objectif de tout individu est de choisir le plus court chemin. (Hypothèses 4). Cela revient à minimiser le coût relatif à chaque chemin.

En se basant sur les différentes hypothèses, et à supposer que le champ de vitesse f, l'inconfort

g et les buts G soient fixes, le choix de chemin dans $\prod$ pour une personne se trouvant à une position x, se fait en minimisant l'expression suivante :

$$Cout = \alpha \int_P 1dS + \beta \int_P 1dt + \gamma \int_P g.dt \tag{2.1}$$

L'équation 2.1 est composée de trois termes relatifs respectivement de gauche à droite, à la longueur des chemins, au temps de parcours et à l'inconfort. On sait par ailleurs que $dS = fdt$. En remplaçant la variable du temps dt par dS dans l'equation 2.1, on obtient :

$$Cout = \int_P \frac{\alpha.f + \beta + \gamma.g}{f} dS \tag{2.2}$$

Ils introduisent la notion de "fonction potentielle" ϕ qui représente en tout point, le coût du chemin optimal vers la destination. Elle est définie comme suit :

$$\begin{cases} \phi(x) = 0, & la\ destination \\ ||\nabla\phi(x)|| = C, & partout\ ailleurs. \end{cases}$$

avec $C = \frac{\alpha.f+\beta+\gamma.g}{f}$. Une meilleure stratégie de déplacement des personnes serait, de toute évidence, un déplacement opposé au gradient de la fonction ϕ. Autrement dit, choisir le chemin avec le coût minimal. Les travaux élaborés en "reconnaissance de comportement de foule" adoptent des approches telles que "l'approche continuum" pour une analyse holistique de la scène. Les approches basées sur l'analyse holistique de la scène sont à leur tour utilisées pour la détection d'événements rares en "vision par ordinateur".

Au niveau microscopique, pour des foules de moyenne densité, on peut retrouver les approches basées agents (agent-based approach) dans lesquelles les individus sont considérés comme des agents autonomes qui peuvent interagir avec leur environnement pour prendre des décisions selon des règles bien établies. Ces modèles sont plus adaptés aux scènes à faible densité où l'on peut facilement détecter et suivre chaque entité. Mais des approches dédiées aux scènes de moyenne densité proposent des adaptations pour leur utilisation. L'un des modèles le plus utilisé dans cette catégorie est le modèle de force sociale, que nous retrouverons plus tard, utilisé également en "vision par ordinateur" pour la détection d'événements rares [52]. Ce modèle a été introduit par Helbing et al. [53] en 1995 et est utilisé pour reproduire certains phénomènes de foule. Il mesure les motivations internes des individus à effectuer un déplacement (ou mouvement) donné. Tout comme le modèle "continuum", le modèle de force sociale se base également sur des hypothèses presque semblables que nous résumons ci-dessous :

- chaque individu cherche à atteindre une destination avec une vitesse voulue, dans une direction voulue,

- le mouvement des individus est influencé par les autres individus présents dans la scène. Il faut également prendre en compte le fait qu'un individu garde toujours une distance minimale par rapport aux autres, mais aussi par rapport aux bâtiments, routes, etc. Cette hypothèse permet de prendre en compte l'évitement des obstacles. Cela est modélisé par une force de répulsion,

- les individus peuvent être attirés par d'autres individus comme les amis, mais également par des rues, etc. Cela se traduit par la modélisation d'une force d'attraction.

Plus tard, Helbing publie dans [54], un travail combinant le modèle de force sociale avec le modèle de panique pour créer un modèle généralisé. Dans ce modèle, la formulation du comportement de la foule intègre les effets psychologique et physique pour être plus réaliste. Le modèle de force sociale se présente comme suit :

$$m_i \frac{dv_i}{dt} = m_i(\frac{v_i^0(t).e_i^0(t) - v_i}{\tau_i}) + F_{int} \tag{2.3}$$

Dans ce modèle, un individu i ayant une masse corporelle m_i, se déplace dans une direction voulue $e_i^0(t)$ avec une vitesse souhaitée $v_i^0(t)$. Il adapte sa vitesse en fonction de la présence ou non d'obstacles. Cette vitesse à un instant t est notée $v_i(t)$. En prenant en compte les interactions que cet individu aura vis-à-vis des autres individus présents dans la foule mais également vis-à-vis des bâtiments qu'il doit éviter, une force d'interaction F_{int} a été rajoutée au modèle comme suit :

$$F_{int} = \sum_{(j \neq i)} f_{ij} + \sum_{w} f_{iw}.$$

avec f_{ij} la force d'interaction entre deux individus i et j et f_{iw} la force d'interaction entre l'individu i et le bâtiment w. Ainsi, l'équation 2.3 modèlise le changement d'allure d'un individu donné. Les détails du calcul des forces d'interaction ainsi que des vitesses peuvent être consultés dans [54].

De noubreux autres travaux [55, 56, 57, 58, 59] ont été et continuent d'être menés sur la thématique. Toutes ces différentes approches visent à modéliser le comportement de foule à des fins surtout de simulations. Les résultats de ces simulations sont généralement utilisés pour la gestion des crises, l'anticipation de phénomènes de foules, l'amélioration des dispositifs de gestion d'évacuation, etc.

2.2.2 Modélisation et analyse de foule en vision par ordinateur

Contrairement aux travaux menés dans le domaine du "computer graphic" [60, 61] qui cherchent également à faire de la simulation ou tendent à créer des animations réalistes des mouvements de foules, les travaux en vision par ordinateur ont pour but d'aller au-delà des simulations et d'offrir des outils d'analyse de scènes pour la détection de mouvements inhabituels à partir de flux vidéo. Ils consistent à construire des algorithmes d'extraction d'informations pertinentes dans une scène et de proposer leur modélisation à l'aide d'algorithmes de "machine learning". Les informations extraites sont de diverses natures et sont parfois issues de l'adaptation des modèles établis en physique.

Mehran et al. dans [52] proposent un modèle de détection de mouvement de foules, plus précisément les mouvements de panique, en se basant sur le modèle de Helbing [55] combiné avec une modélisation en "sac de mots". Le modèle suit les étapes classiques des systèmes de classification connus en "vision par ordinateur" : **Extraction de caractéristiques**, suivi de la **Construction du modèle** et puis de la **Classification** (voir Fig 2.2). Au niveau de l'extraction des caractéristiques, ils proposent de calculer la **force sociale** des individus présents dans la scène à partir d'un modèle légèrement modifié, intégrant un "poids de panique" et prenant en compte

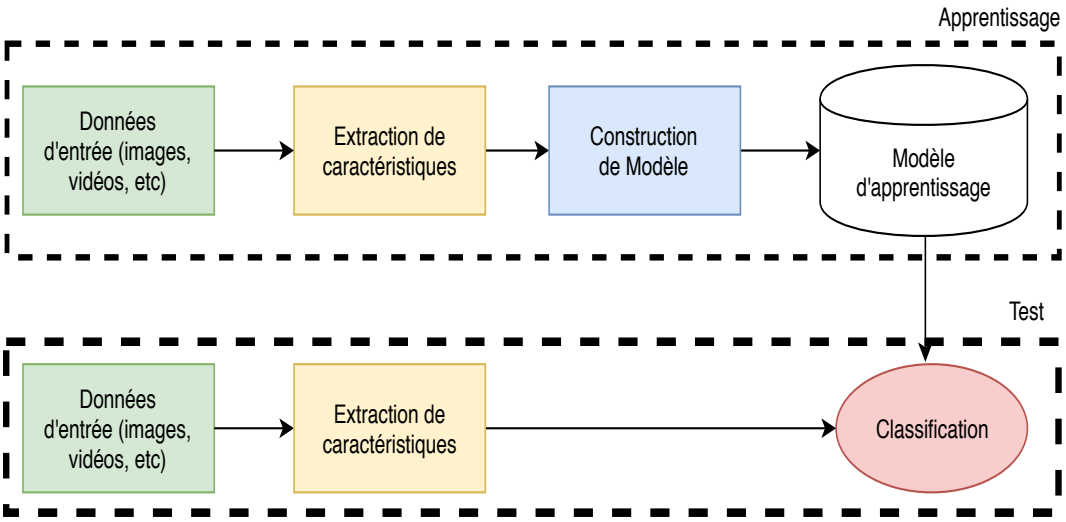


FIGURE 2.2 – Schéma classique d'apprentissage et de classification

l'influence de la vitesse du voisinage. Ainsi, le modèle proposé revient à :

$$m_i \frac{dv_i}{dt} = m_i(\frac{v_i^q - v_i}{\tau_i}) + F_{int} \tag{2.4}$$

avec $v_i^q = (1 - p_i) * v_i^p + p_i * < v_i^c >$.

p_i est le "poids de panique" attribué à l'individu, v_i^p la vitesse désirée et $< v_i^c >$ la vitesse moyenne dans son entourage. Ce modèle suppose une détection et un suivi de chaque individu dans la scène, ce qui est impossible pour une scène de moyenne ou forte densité. Pour résoudre ce problème, Mehran et al. proposent de disposer sur une grille régulière posée sur l'image, des particules qui seront déplacées à l'aide du flot optique calculé dans la vidéo. Pour chaque particule se trouvant à la position (x_i, y_i), la vitesse moyenne de l'entourage $< v_i^c >$ est la moyenne du flot optique dans une petite fenêtre centrée en ce point $< v_i^c >= O_{ave}(x_i, y_i)$ et la vitesse désirée $v_i^p = O(x_i, y_i)$ est la valeur du flot optique en ce point. Le calcul de la force sociale qui n'est rien d'autre que la force d'interaction, peut être réalisé à partir de l'équation 2.4. Une fois le calcul des forces d'interaction entre particules effectué, ils génèrent des cartes qui correspondent à l'amplitude du vecteur de la force d'interaction en chaque point. La modélisation des événements a été réalisée avec l'algorithme LDA (Latent Dirichlet Allocation), qui prend en entrée une représentation sous forme de document des zones spatio-temporelles, dans un ensemble d'images (clips), n'ayant pas une force d'amplitude nulle. La classification des clips en normal ou anormal est faite en fixant un seuil. Les résultats de l'évaluation montrent que le modèle de force sociale permet de bien capter la dynamique de la scène et permet de détecter les événements relatifs à la panique de foule. Ultérieurement à cette méthode, Mehran et al. introduisent dans [62] le concept de "fonctions potentielles" composées de deux parties : une fonction de flux et une fonction de vitesse. La fonction de flux donne des informations sur la partie stable et non divergente du flux et la fonction de vitesse donne les informations sur les changements locaux dans les mouvements. Ce concept est inspiré de la méthode de Lagrange pour la dynamique de fluide et a été utilisé pour d'une part segmenter les mouvements de foule et d'autre part pour la détection d'événements

anormaux. La méthode consiste à utiliser une séquence d'images pour générer des cartes à partir de la "fonction potentielle". Ces cartes sont utilisées comme des données d'entrée de l'algorithme SVM pour la classification des images.

Benabbas et al. [63] proposent d'utiliser le flot optique pour faire une modélisation de l'information du mouvement global dans une scène. A partir des informations d'orientation et d'amplitude, ils modélisent les mouvements dominants dans des blocs de la scène (la scène est divisée en des blocs de tailles égales) en créant des modèles de direction et d'amplitude. En utilisant un algorithme de segmentation de région, ils proposent de regrouper ensemble les blocs voisins qui sont similaires en terme de vitesse et de direction. La détection des événements est réalisée en analysant les flots optiques à l'intérieur des groupes de blocs formés et en les comparant à des modèles appris. Il est à noter que dans ce papier, différentes catégories de mouvements sont considérées. Ainsi, ils distinguent trois catégories de mouvements à savoir : les mouvements relatifs à la vitesse, les mouvements de convergence de foule et les mouvements de divergence de foule. Le type de mouvement détecté est fonction de l'interaction entre les groupes formés précédemment. Par exemple, si deux ou plusieurs groupes se rapprochent l'un de l'autre, il s'agit d'un événement de convergence de foule. Les limites de la méthode résident justement dans cette catégorisation des événements à détecter car cela suppose une connaissance à priori de ces derniers.

Tian wang et al. [64] proposent une représentation sous forme d'histogramme du flot optique, calculé dans des blocs réguliers de l'image et leur modélisation à partir d'un algorithme de "SVM One class". Cong et al. [65] se basent sur les histogrammes multi-échelles (MHOF) à 16 cases (bins) du flot optique, calculés dans des unités de bases de forme 2D ou 3D (spatio-temporel). Ensuite, les MHOF issus de toutes les unités de base de toute l'image sont concatenés pour former un seul vecteur de caractéristiques. Dans l'ensemble des vecteurs de caractéristiques du jeu d'apprentissage, un algorithme de sélection de sous-ensemble optimal permet d'avoir un jeu de données de taille réduite mais représentatif du jeu de données, pour former un dictionnaire. Le dictionnaire ainsi formé, combiné avec une matrice de poids, permet de facilement reconstruire tout le jeu d'apprentissage de départ. La détection d'événements anormaux, consiste à calculer le coût de la reconstruction (Sparse Reconstruction Cost) du vecteur de caractéristiques d'une séquence d'images test à partir du dictionnaire d'apprentissage et de la matrice de poids associée. L'expérience montre que le côut de reconstruction pour des séquences d'images contenant des événements normaux est très bas comparativement à celui pour des images contenant des événements inhabituels. Un seuillage permet facilement de classer les séquences d'images en normales ou anormales.

Yinghuan Shi et al. [66] proposent une méthode qui utilise la correlation de phase, basée sur le théorème de décalage de Fourier, pour estimer le vecteur de mouvement entre deux images successives. Les images sont divisées en des régions adjacentes pour lesquelles les vecteurs de mouvements sont extraits. L'algorithme STCOG (Spatial-Temporal Co-Occurence Gaussian) est utilisé pour estimer la probabilité de co-occurence entre des vecteurs de mouvements de régions locales. Cette probabilité est utilisée pour décider si une région contient des événements anormaux

ou non. Leur méthode nécessite juste une courte période d'entraînement et présente une faible complexité de calcul.

Plus récemment, Singh et al. [67] ont proposé une approche mixte pour les détections globales et locales. Leur méthode consiste à faire une modélisation sous forme de graphe de la scène, en utilisant des points d'intérêt spatio-temporels détectés dans la scène et qui servent de sommets du graphe ainsi que la description de ces points d'intérêt qui sert à établir les liens entre ces derniers. Pour la détection globale, ils adoptent une approche "sac-de-mots" version graphe appelée "sac-de-graphes" qui reprend les étapes traditionnelles de l'approche "sac-de-mots" avec des algorithmes de clustering adaptés aux graphes. La classification est faite avec un "SVM binaire".

Fang et al. [68] et Bao et al. [69] ont proposé des méthodes similaires basées sur l'utilisation de l'agorithme PCANet pour l'extraction de caractéristiques de haut niveau. La démarche consiste à extraire des caractéristiques à un niveau supérieur, à partir d'une série successive d'analyse en composantes principales effecutées sur des caractéristiques intermédiaires vues avec les approches précédentes. Fang et al. utilisent l'histogramme des orientations du flot optique combiné avec l'information de saillance comme entrée de l'algorithme de PCANet. Les vecteurs de caractéristiques issus de la dernière couche du PCANet sont utilisés pour construire un modèle avec un SVM. Bao et al. quant à eux, utilisent directement l'information du flot optique comme entrée du PCANet. Les vecteurs issus de la dernière couche sont utilisés pour créer des groupes. La classification est faite en utilisant la distance entre le vecteur de test et tous les centroïdes des groupes précédemment formés. Les deux méthodes se situent à la frontière des nouvelles méthodes développées en "machine learning" et qui sont basées sur des réseaux convolutifs.

Tableau 2.1 – Tableau récapitulatif des méthodes de détection globale en vision par ordinateur

Référence	**Caractéristiques**	**Algo d'apprentissage**	**Dataset**
Mehran et al. [52]	Force Sociale	LDA	UMN
Yanghuan et al. [66]	Correlation de Phase	STCOG	UMN
Benabbas et al. [63]	Flot Optique (FO)	-	PETS/CUHK
Cong et al. [65]	MHOF	SRC	UMN
Tian et al. [64]	HOFO	SVM one class	UMN/PETS
Fang et al. [68]	MHOF+SI+PCANet	SVM one class	UMN
Bao et al. [69]	FO + PCANet	Distance Centroïd	UMN
Singh et al. [67]	HOG + HOF + graphe	SVM	UMN

Au regard des nombreuses approches proposées, on se rend compte que la détection d'évènements rares ou anormaux pour des foules de moyenne ou haute densité, a toujours suscité et continue de susciter l'intérêt des chercheurs dans la communauté de vision par ordinateur. Le tableau 2.1 récapitule les méthodes qui font la détection globale pour des foules de moyenne ou forte densité et que nous avons présenté ci-dessus. Notre méthode s'inscrit dans le schéma clas-

sique de classification et propose une extraction de caractéristiques dans des régions sélectionnées pour leur saillance visuelle et la modélisation à travers l'approche "sac-de-mots". Le détail de la méthode ainsi que les résultats obtenus seront présentés au chapitre 3.

2.3 Détection et localisation d'événements rares locaux dans une scène à faible densité de foule

Outre les approches dédiées à la détection d'événements anormaux dans une scène à moyenne ou forte densité de foule, d'autres approches se sont consacrées à la détection et à la localisation dans un environnement à faible densité de personnes. A priori, pour des scènes à faible densité, la difficulté de modélisation doit être moindre comparée aux autres scènes. Néanmoins, d'autres difficultés liées à la localisation des événements émergent. Les approches présentées ci-dessus échouent dès lors que la densité des événements inhabituels est largement inférieure à la densité des événements normaux dans la scène. Prenons le cas de la surveillance du trafic routier. Quand un motocycliste tombe sur la voie et que dans le même temps, les autres usagers continuent de circuler en évitant juste celui qui est tombé, la modélisation globale de la scène peut s'avérer inefficace. Une extraction locale des informations et une modélisation complexe en vue d'inclure toutes les différences intra-classes des événements normaux constituent l'une des solutions devenues classiques dans la littérature. Comme pour les scènes à moyenne ou forte densité, deux classes d'approches se dégagent : les approches inspirées de la physique et les approches purement "vision par ordinateur".

2.3.1 Les approches inspirées de la Physique

Tout comme les méthodes décrites ci-dessus dans le cadre de la détection globale d'événements rares et qui se sont inspirées de la physique, les méhodes qui s'attèlent à localiser les événements dans la scène, utilisent presque les mêmes modèles ou algorithmes de la physique combinés avec des algorithmes de "machine learning". On peut identifier trois différentes catégoires d'approches de la physique abondamment utilisées dont certaines ont été déjà détaillées plus haut.

2.3.1.1 Les modèles de champ d'écoulement

Les modèles de champ d'écoulement permettent de suivre et de comprendre l'évolution de la foule dans le temps. Tout changement de mouvement peut alors être observé et surtout localisé. Wu et al. [70] ont proposé une méthode pour à la fois des scènes structurées et non-structurées, basée sur l'extraction de la trajectoire de la foule à partir de particules déposées sur l'image et qui sont déplacées par le flot optique. Ils proposent par la suite de regrouper les trajectoires en des clusters qui serviront à l'extraction de la "dynamique chaotique" de la foule à l'aide des "invariants chaotiques" connus sous le nom de "maximal Lyapunov exponent and correlation dimension". Ces caractéristiques sont utilisées pour apprendre un modèle probabiliste sur lequel le critère du maximum de vraissemblance est appliqué pour classer les séquences d'images en normales ou anormales. La localisation dans la scène est faite en faisant un clustering sur les

positions des différents vecteurs de caractéristiques issus des images classées anormales. Les clusters avec peu de trajectoires sont supprimés pour laisser place au reste qui est considéré comme contenant les trajectoires des régions d'anormalité. Cette manière de localisation n'est pas tout à fait efficace puisque la détection de l'anormalité est d'abord globale. La méthode ne peut fonctionner que pour des cas de détection précis.

Chen et al. [71] proposent une méthode de détection des mouvements saillants dans une scène de foule de densité variée. La méthode consiste à créer des cartes de saillances à partir de l'angle de phase et de l'amplitude du flot optique en chaque point de la scène, en faisant une analyse spectrale indépendante de chaque information du flot. L'analyse spectrale consiste à faire une transformation dans le domaine de Fourier des images des orientations et des phases et puis de calculer le résidu spectral de chaque image. Une transformation inverse dans le domaine spatial, permet de mettre en évidence les régions avec d'une part des orientations qui diffèrent de l'ensemble des orientations de la scène et d'autres part les régions avec des amplitudes particulières. Une combinaison des deux cartes de saillances est faite en prenant pour chaque pixel, le maximum entre les valeurs de saillances des deux cartes. La méthode ne nécessite aucune phase d'entrainement avant de détecter les régions contenant des actions saillantes. Cette approche, bien qu'intéressante, présentera des difficultés pour une détection et la localisation d'événements rares dans certaines scènes surtout dans un contexte non structuré. On peut dire qu'il ne s'agit pas à proprement dit, d'une méthode de détection d'événements rares, puisque les actions saillantes ne sont pas forcémment des actions qui peuvent générer des événements rares. Il s'agit plutôt d'une méthode de segmentation d'actions dans une scène. Celle-ci peut alors être combinée avec d'autres méthodes afin d'obtenir de meilleurs résultats.

2.3.1.2 Les modèles de forces sociales

Les modèles de forces sociales, que nous avons présentés dans la section précédente, ont été également utilisés pour la détection locale d'événements. Mehran et al. [52] en introduisant la force sociale dans la détection d'événements rares dans une foule, s'étaient arrêtés à une détection globale. Dans cette approche, ils suggèrent que les régions où l'amplitude de la force sociale est élevée peuvent être considérées comme les localisations des événements rares. Cette supposition n'est valable que pour des scènes dont la densité de l'anormalité est grande.

Raghavendra et al. [72] s'inspirent de la méthode de Mehran et al. [52] et proposent une extension de cette dernière. Dans cette version étendue de l'approche, ils proposent d'optimiser la force sociale en utilisant la méthode PSO (Particle Swarm Optimization) pour le déplacement des particules sur l'image. Le but est de déplacer les particules vers les zones principales de mouvement de l'image dans le but de minimiser la force d'interaction afin de modéliser les comportements les plus diffus et normaux de la foule. Les anomalies peuvent être détectées en vérifiant si certaines particules (forces) ne correspondent pas à la distribution estimée, et ceci par une méthode de type RANSAC suivie d'un algorithme de segmentation pour localiser finement les zones anormales. Zhao et al. [73] proposent d'utiliser la notion d'instabilité de foule pour la détection des événements rares. Ils introduisent une probabilité de collision dans le modèle de

force sociale pour diminuer les fausses alarmes qu'engendraient l'utilisation de la SF dans les travaux précédents. Pour chaque particule dans la scène, ils estiment le flot optique ansi que la force sociale. Afin de réaliser la détection et la localisation des événements, la scène est subdivisée en blocs et pour chaque bloc, le vecteur de caractéristiques est composé de la moyenne du flot optique et de la force sociale. Ces vecteurs sont utilisés pour faire une évaluation de la distribution spatiale de l'instabilité de la foule. Enfin, l'instabilité temporelle est déterminée à partir d'une analyse statistique de l'instabilité de tous les blocs.

2.3.1.3 Les modèles d'énergie de foule

Il a été également exploité d'autres types de caractéristiques pour détecter et localiser les événements. La dernière catégorie que nous présentons est basée sur l'estimation de l'énergie d'une foule à partir de différentes méthodes. Ainsi, Yang et al. [74] ont proposé d'exploiter la pression locale dans une foule pour la détection d'anormalité. Le calcul de la pression est fait en combinant le modèle de force sociale et la densité locale extraite avec l'algorithme de LBP (Local Binary Pattern). Une représentation en forme d'histogramme de la direction de la pression (HOP) est utilisée pour entraîner un modèle SVM. L'évènement est localisé au niveau des particules avec une forte densité locale.

Ren et al. [75] proposent d'utiliser l'entropie des changements de comportement dans la scène pour détecter et localiser les anormalités. En utilisant les modèles de la théorie de l'information et en se basant sur le flot optique des pixels ayant une amplitude non nulle, ils calculent l'entropie de ces pixels. Les images contenant des événements rares sont celles ayant une entropie supérieure à un seuil donné. Pour la localisation, ils considèrent juste les régions avec de fortes valeurs d'entropie dans l'image.

Tableau 2.2 – Tableau récapitulatif des méthodes de détection locale inspirées de la physique

Référence	**Caractéristiques**	**Algo d'apprentissage**	**Dataset**
Mehran et al. [52]	Force Sociale	LDA	UMN
Wu et al. [70]	Invariance chaotique	-	UMN
Chen et al. [71]	Saillance de Flot	-	-
Raghavendra et al. [72]	FO + SFM-SPO	RANSAC	UCSD
Zhao et al. [73]	champ de vitesse + SFM	-	UMN
Yang et al. [74]	Modèle de pression	SVM	UMN
Ren et al. [75]	Entropie	-	UMN/UCSD

Le tableau 2.2 résume les méthodes utilisées en "vision par ordinateur" et inspirées des modèles de la physique pour la détection et la localisation des événements rares. Nous aborderons dans la section suivante, les méthodes qui sont purement issues de la communauté de "vision par ordinateur" sans inspiration des modèles de la physique.

2.3.2 Les approches purement vision par ordinateur

L'évolution des recherches en "traitement d'image", "vision par ordinateur" et en "machine learning" a fait progresser les travaux sur la détection et la localisation des événements rares dans une scène. On retrouve ainsi plusieurs méthodes, pour la plupart récentes, qui exploitent différentes caractéristiques visuelles et ont recours à des algorithmes avancés en "machine learning" pour leur modélisation.

Kim et al. dans [76], abordent le problème en introduisant une représentation des vidéos à l'aide du modèle MRF ("Markov Random Field") spatio-temporel. Cette représentation consiste à diviser les vidéos en des grilles de régions spatio-temporelles, qui constituent les noeuds du MRF, et à connecter les régions voisines par des liens. Chaque noeud est associé à des observations continues de flot optique. Ils servent à modéliser des activités via une analyse en composante principale basée sur l'algorithme MPPCA ("Mixture of Probabilistic Principal Component Analysers"). Sur la base des modèles appris, les paramètres du MRF sont estimés. Une inférence bayésienne sur le MRF permet d'obtenir une estimation probabiliste pour déterminer les noeuds qui sont normaux ou non. Une mise à jour incrémentale du MPPCA et du MRF offre la possibilité d'adaptation du modèle à de nouveaux jeux de données. La méthode a été testée sur des vidéos de surveillance de portes d'entrée et de sortie d'un lieu donné et présente de bons résultats. Néanmoins, le test sur des bases de vidéos avec des événements plus complexes, montrent la faiblesse de la méthode.

Kratz et al. [77] se sont basés sur l'un des algorithmes les plus utilisés en modélisation (les modèles de Markov cachés) pour construire des modèles locaux d'événements, par zone de la scène. Les modèles de Markov cachés permettent de capter la nature intrinsèquement dynamique des caractéristiques observées. Le gradient spatio-temporel de chaque pixel d'un volume d'images (appelé cuboïd) est calculé pour servir de caractéristiques visuelles à la modélisation. Ils proposent de capter la relation temporelle entre les modèles de mouvements locaux des cuboïds d'une même région à travers de simples modèles de Markov caché et d'utiliser des modèles de "Markov cachés couplés" entre les régions voisines pour capter la relation spatiale qui les lie. Les événements rares sont considérés comme étant des déviations statistiques par rapport à ses modèles appris. Leur méthode a prouvé son efficacité pour la localisation dans des scènes de forte densité. Néanmoins, l'utilisation d'un seul modèle de Markov par région constitue une faiblesse de la méthode car elle est limitée à un nombre donné d'événements à modéliser et nécessiterait un ré-entrainement au cas où le type d'événement changerait. Wang et al. [78] ont également fait appel aux modèles de Markov cachés pour les mêmes tâches. A la différence de Kratz et al. [77], ils proposent d'utiliser en plus du gradient spatio-temporel, des transformations d'ondelettes pour extraire les informations de hautes fréquences des cuboïds. De multiples modèles de Markov cachés sont employés pour faire la modélisation et chaque modèle compte pour un type de comportement.

Suite à la mise en place du modèle MDT (Mixture Dynamic Texture) par Chan et al. [79] pour la segmentation de mouvements, certains travaux l'ont adapté pour la détection et la loca-

lisation d'événements rares. Le MDT est un modèle génératif spatio-temporel, qui représente des séquences vidéos comme des observations, à partir des variations dynamiques linéaires, pour en extraire des propriétés stationnaires spatio-temporelles. Mahadevan et al. [80, 81] proposent une modélisation locale par région des événements en utlisant le MDT. La classification de chaque région est faite en calculant la log probabilité négative par rapport au modèle de la région. Le MDT capte à la fois l'information d'apparence et celle de mouvement dans la scène. Leur méthode présente de meilleurs résultats comparativement aux anciennes méthodes basées uniquement sur le flot optique.

Cong et al. [65] ont démontré que leur méthode basée sur les histogrammes multi-échelles du flot optique (MHOF) et du coût de reconstruction (Sparse Reconstruction Cost) à partir d'un dictionnaire appris, s'adapte bien à ce nouveau défi. La différence avec la précédente méthode, se situe dans l'extraction des descripteurs MHOF, qui à défaut de se faire sur toute l'image, se fait dans des zones ayant des structures de différentes formes (étoiles, cuboid, etc.). Chaque structure est composée de plusieurs unités de base. Les MHOF extraits dans les unités de bases qui constituent ces structures, sont concatenés pour donner un vecteur de caractéristiques unique par structure. Le reste de la méthode reste inchangé. Profitant de la performance du SRC et du MDT, Xu et al. [82] proposent une méthode combinant les deux précédentes méthodes. Elle consiste à faire la reconstruction éparse de la dynamique de la texture qui est décrite ici par la méthode LBPTOP (Local Binary Patterns from Three Orthogonal Planes). Ainsi, une région est déclarée contenir un événement inhabituel, si l'erreur de la reconstruction de la dynamique de la texture est élevée.

Un autre groupe de méthodes utilise l'approche "sac-de-mots". Cette approche se base sur l'extraction de caractéristiques visuelles bas niveau (le mouvement ou l'apparence) pour créer des sous-groupes sur l'ensemble des données d'apprentissage. La méthode de Mehran et al.[52] bien que faisant partie de cette catégorie, n'est pas adaptée à la localisation des événements. Wang et al. [83] ont également utilisé l'approche "sac-de-mots" avec des données spatio-temporelles extraites dans des cuboïds. En introduisant une nouvelle structure de données, ils ont rendu possible le calcul de similarité entre des cuboïds de tailles différentes. L'algorithme LDA est utilisé par la suite pour la modélisation des événements. Dans la même catégorie, Roshtkhari et al. [84] proposent une extension de l'approche "sac-de-mots" pour la détection et la localisation d'événements suspicieux. Cette approche ne fait appel à aucun descripteur et n'utilise comme jeu d'apprentissage, que la vidéo elle-même. Autrement dit, c'est une méthode qui apprend au fur et à mesure des nouvelles observations et qui se base uniquement sur les valeurs des pixels. La méthode commence par une subdivision de la vidéo en des volumes appelés "composition spatio-temporelle". Dans un second temps, un codebook est formé à partir de l'ensemble des compositions afin de regrouper les compositions similaires et de réduire les informations redondantes. Puis, les relations de compositions spatio-temporelles entre les volumes du même cluster sont approximées à partir d'un modèle de mélange gaussien. Pour détecter et localiser les événements, ils proposent de construire un ensemble de volumes, d'assigner à chaque volume un cluster

et de calculer la probabilité de l'ensemble de volumes au regard de la fonction de densité estimée. La décision de classer les volumes est prise en fixant un seuil sur la probabilité. La méthode nécessite néanmoins une phase d'initialisation qui est faite à partir d'un nombre minimum de séquences de vidéos contenant uniquement des événements normaux.

Dan et al. [85] proposent une méthode mixte de détection globale et locale. La méthode présente deux niveaux d'extraction d'informations qui servent à la détection des évènements anormaux. Il s'agit de subdiviser la scène en de petites cellules (idéalement 10×10 px) et d'en extraire l'histogramme des flots optiques. La modélisation des évènements est effectuée en plusieurs étapes. Dans un premier temps, les différents vecteurs de flots sont regroupés en k groupes qui représentent des activités atomiques. Ces activités sont, par la suite, modélisées en déterminant le vecteur de distribution des clusters qui est obtenu en calculant la distance au centre des points appartenant au cluster. A partir de cette distribution et pour chaque vecteur de caractéristiques, la densité de probabilité par rapport aux clusters est déterminée en utilisant une fonction fenêtre de Parzen. Le vecteur obtenu à la fin de cette étape (de taille k) est appelé histogramme d'activités. Dans un second temps, ils ont proposé de modéliser les évènements par régions pour permettre la localisation. Chaque région de la scène étant représentée par n cellules obtenues des n images d'apprentissage, un algorithme permet de déterminer les m plus pertinents vecteurs d'activités tout en considérant les vecteurs des 4 voisins adjacents de chaque cellule. A la fin de cette étape, chaque region est représentée par m vecteurs d'activités. La classification des différentes régions dans une vidéo de test, est réalisée grâce à une fonction d'énergie proposée par les auteurs.

Shi et al. [86] proposent une méthode basée sur la saillance visuelle. Le premier problème abordé dans leur travail est celui relatif à la profondeur de vue, c'est-à-dire le changement de taille des objets en fonction de leur position dans la scène. Ils proposent une modélisation par région, contrairement à la modélisation par bloc, en divisant la scène en des régions. Les régions sont des parties de la scène dont les objets et personnes ont des tailles similaires. Les informations exploitées pour la modélisation sont le flot optique et la saillance visuelle. Dans une première étape, la carte de saillance spatiale et celle de l'énergie du flot optique sont calculées. Une fonction de fusion qui consiste à multiplier élément à élément les deux cartes, est utilisée pour obtenir une carte finale de saillance spatio-temporelle. La construction des modèles de chaque région est faite à partir de l'histogramme multi-échelles de l'amplitude du flot optique extrait dans des unités de bases formées des blocs sélectionnés en fonction de leur score de saillance. L'algorithme KNFST (Kernel Null Foley-Sammon Transform) est utilisé pour faire la modélisation.

La méthode de Singh et al. [67] qui modélise les événements en utilisant les graphes et que nous avons mentionnée ci-dessus, présente également de bonnes performances dans la détection locale. Dans sa version adaptée à la localisation, les auteurs ont eu recours à l'algorithme SVM avec un noyau adapté aux graphes. La classification a lieu sur chacun des graphes formés dans la vidéo. Un graphe classé anormal signifie que les zones de la scène qui contiennent les points d'intérêt qui constituent les sommets du graphe sont les emplacements de l'événement anormal.

Tableau 2.3 – Tableau récapitulatif des méthodes de détection locale issue de la vision par ordinateur

Référence	**Caractéristiques**	**Algo d'apprentissage**	**Dataset**
Kim et al. [76]	MRF + MPPCA	Inférence bayésienne	-
Kratz et al. [77]	Gradient 3D	Markov	-
Wang et al. [78]	G3D + wavelet	Markov multiple	youtube
Mahadevan et al. [80][81]	MDT	-	UCSD
Cong et al. [65]	MHOF	SRC	UCSD/Subway
Xu et al. [82]	LBPTOP	SRC	UCSD/PETS
Dan et al. [85]	HOF+Histo Activité	Fonction d'Energie	UCSD
Shi et al. [86]	FO + Saillance	KNFST	UCSD
Singh et al. [67]	HOG+HOF+graphe	SVM	UCSD
Roshtkhari et al. [84]	STC	-	UCSD/Subway

L'ensemble des méthodes présentées ci-dessus et résumées dans la tableau 2.3 utilisent toutes des caractéristiques extraites à la main (handcraft en anglais). Une nouvelle génération de méthodes basées sur les réseaux convolutifs a vu le jour ces dernières années. Petit à petit, elles sont utilisées en analyse de scène. Dans la section suivante, nous présenterons quelques travaux récemment menés et utilisant les réseaux convolutifs.

2.4 La nouvelle génération d'approches basées sur le deep learning

Depuis les années 2012, la communauté de "vision par ordinateur" a connu un boulersement important avec le développement des méthodes basées sur les réseaux de neurones convolutifs. L'idée de l'utilisation des réseaux de neurones ne date pourtant pas des années 2012 mais bien avant, vers les années 1998 avec le célèbre papier de Yann LeCun intitulé "Gradient-based learning applied to document recognition" [87] dans lequel il proposa l'entraînement d'un réseau convolutif avec l'utilisation de la méthode de "back-propagation". Mais à partir de 2012, avec la performance extraordinaire du réseau AlexNet, proposé par Alex et al. [88] au challenge "ImageNet Large-Scale Visual Recognition Challenge", le nombre de travaux utilisant les méthodes dites "deep learning" (apprentissage profond en français) a explosé. Tous les domaines d'application semblent touchés même si certains checheurs se montrent réticents en la considérant comme une boîte noire. Ainsi, récemment, des travaux en détection d'événements anormaux basés sur l'utilisation des approches "deep learning" ont vu le jour. Ces méthodes s'appuient sur les techniques d'apprentissage de caractéristiques. En effet, en "machine learning", les approches dites "d'apprentissage de caractéristiques" ou plus généralement "representation learning" découvrent automatiquement à partir de données bruitées, des représentations ou caractéristiques adéquates pour les tâches de classification ou de détection. Le but premier de ces approches est de s'af-

franchir de l'étape d'extraction manuelle de caractéristiques, jusqu'ici utilisée dans les systèmes de classification, en les apprenant automatiquement. Autrement dit, il s'agira de construire des modèles paramétrés, qui reconstruisent les données d'entrée, $f_\theta : \mathcal{X} \rightarrow \mathcal{Z} \rightarrow \mathcal{X}$ tel que l'espace latent $\mathcal{Z}$ soit invariant aux changements de luminosité et de translation dans les données d'entrée. Dans cette section, nous présenterons brièvement trois catégories d'approches "deep learning" utilisées pour résoudre le problème de la détection d'événements. Une revue complète des méthodes utilisant le "deep learning" peut être consultée dans [89].

2.4.1 Les Modèles de reconstruction de données

La reconstruction de données consiste à créer de nouvelles données à partir d'un modèle appris. Dans cette catégorie d'approches, il sera question d'apprendre une représentation optimale des données à partir de l'ensemble du jeu d'apprentissage ne contenant que des événements normaux pour, par la suite, les reconstruire. L'hypothèse qui sous-tend cette démarche est que l'erreur de reconstruction des données de départ à partir du modèle de représentation appris doit être petite et qu'à l'opposé, l'erreur de reconstruction d'une donnée non apprise serait grande. On retrouve dans cette catégorie, les modèles d'analyse en composante principale (ACP) ainsi que les différentes formes d'AutoEncodeurs (AEs).

2.4.1.1 Analyse en Composante Principale

L'ACP modélise la correlation spatiale entre les pixels d'une image. A travers cette modélisation, l'objectif de la méthode ACP est de trouver un ensemble de projections orthogonales qui décorrélent les caractéristiques de l'ensemble d'apprentissage. Les images étant vectorisées, le jeu d'apprentissage revient à une matrice centrée $X \in \mathbb{R}^{N \times d}$ avec N le nombre d'images de la base et $d = h \times w$ la dimension du vecteur. Dans l'équation 2.5, une réduction de la dimensionnalité des données d'entrée à travers XW est faite avant la reconstruction qui consiste à reprojecter les données réduites dans l'espace initial. Les anormalités sont captées par le calcul de la distance entre la donnée initiale et la donnée reconstruite. Dans le cadre de la détection d'événements rares, les images de flot optique peuvent être utilisées comme données d'entrées d'une telle approche.

$$\min_{W^T W = I} ||X - (XW)W^T||_F^2 = ||X - \hat{X}||_F^2 \tag{2.5}$$

avec $W \in \mathbb{R}^{N \times k}$ (matrice de projection dans un sous-espace réduit) et sous la contrainte $W^T W = I$.

2.4.1.2 Les Auto-Encodeurs

Les Auto-encodeurs sont une alternative aux approches ACP pour la réduction de dimensionnalité. Il s'agit d'un réseau de neurones entraîné avec la méthode de "back-propagation" par réduction des erreurs de reconstruction du jeu d'entrée. Très semblables au perceptron multicouches, les auto-encodeurs ont des couches d'entrée et de sortie ainsi qu'une ou plusieurs autres couches cachées qui relient celles d'entrée et de sortie (voir fig 2.3). Les couches d'entrée et de sortie possèdent le même nombre de noeuds, étant donné que le but est de reconstruire le jeu

d'entrée après la réduction de dimension. Ainsi, les auto-encodeurs sont constitués de deux parties à savoir : l'encodeur, qui prend les données d'entrée et les projecte dans le sous-espace de dimension inférieure (espace latent) et le décodeur qui a pour but de reconstruire l'entrée à partir des données du sous-espace latent. La projection dans l'espace latent est une transformation non-linéaire utilisant des fonctions d'activation :

$$z = \sigma(Wx + b)$$

avec $x \in \mathbb{R}^d$ la donnée d'entrée, $z \in \mathbb{R}^k$ la variable latente (encore appelée code), W la matrice de poids et b un vecteur de biais. La fonction σ est généralement soit un ReLU (Rectified Linear Unit) $(\sigma(x) = max(0, x))$ soit un sigmoïd $(\sigma(x) = (1 + e^{-x})^{-1})$.
La recontruction de l'entrée, donc le décodage, possède également la même expression que l'encodage et peut utiliser les mêmes paramètres (c'est à dire la matrice de poids et le vecteur de biais) :

$$x' = \sigma'(W'z + b')$$

L'expression de l'évaluation de l'erreur de reconstruction est similaire à celle de l'équation 2.5 :

$$\min_{U,V} ||X - \sigma(XU)V||_F^2 \tag{2.6}$$

avec U, V les paramètres de régularisation à optimiser pour apprendre une représentation optimale du jeu d'entrée. D'autres variantes d'auto-encodeurs tels que les auto-encodeurs convolu-

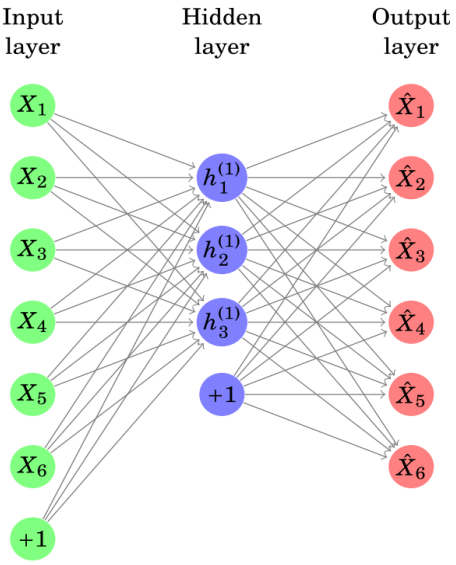


FIGURE 2.3 – Exemple d'Auto-encodeurs [89]

tionnels, les auto-encodeurs épars, les auto-encodeurs contractifs, les auto-encodeurs débruiteurs, ont été élaborés pour améliorer la version initiale.

Dans [90], Sabokrou et al. ont utilisé deux types d'auto-encodeurs, un épars et un non épars. La méthode consiste à extraire des cuboïds dans une grille régulière sur l'image. Les cuboïds de taille 10 x 10 x 5 sont utilisés en entrée de l'auto-encodeur épars qui évalue leur "sparsity". Les cuboïds dont la représentation n'est pas assez épars, sont utilisés comme points d'intérêt autour

desquels de nouveaux cuboïds de taille 30 x 30 x 10 sont extraits et serviront d'entrée au second auto-encodeur non-épars qui évaluera leur erreur de reconstruction. Ainsi, les cuboïds qui ne sont pas assez épars à petite échelle et dont l'erreur de reconstruction à grande échelle est grande, sont considérés comme contenant des événements anormaux.

2.4.2 Les Modèles prédictifs

Comme son nom l'indique, un modèle prédictif a pour but de prédire l'image à un instant t. Cette prédiction est faite en se basant sur de précédentes observations. Autrement dit, il s'agit de modéliser la distribution conditionnelle $P(x_t|(x_{t-1}, x_{t-2}, ..., x_{t-p}))$ à partir d'un petit ensemble de données observées. Les modèles les plus utilisés dans cette catégorie sont les modèles auto-regresseurs et les modèles LSTM (Long Short-Term Memory).

2.4.2.1 Le Modèle "Long Short-Term Memory"

Les LSTMs sont une forme particulière des RNN (Recurrent Neural Network) qui ont été utilisés abondamment dans la résolution de plusieurs tâches telles que la reconnaissance de la parole, la traduction, la modélisation du langage, etc. Les RNN sont des réseaux de neurones avec boucles qui permettent à l'information de persister dans le réseau. Ils peuvent également être vu comme un ensemble de réseaux successifs dont le résultat au niveau d'un noeud dépend de l'information de sortie envoyée par le noeud précédent (voir Fig 2.4). Le désir d'exploiter au

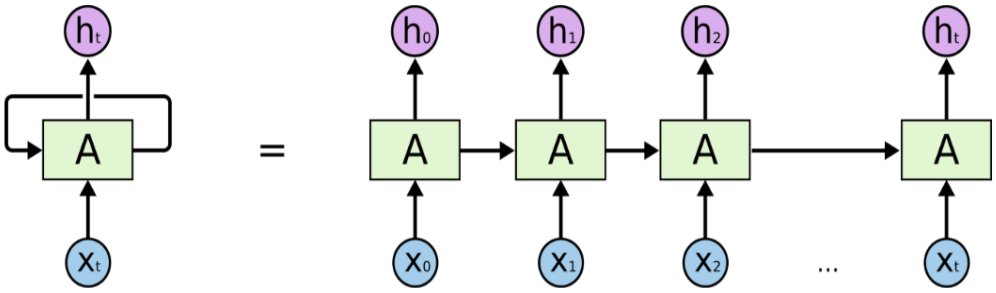


FIGURE 2.4 – Architecture d'un RNN : A gauche un RNN et à droite sous sa forme déroulée

maximum les informations passées révèle rapidement les limites des RNNs dues aux problèmes de "dépendances à long terme" qui s'observent avec les tâches pour lesquelles la prise en compte du contexte est importante pour l'apprentissage. Un vrai écart s'observe entre le moment de l'apprentissage de l'information pertinente et le moment de son utilisation. Et plus cet écart augmente, plus les RNNs ont du mal à apprendre à connecter les informations disponibles.

Les LSTMs ont été proposés par Hochreiter et Schmidhuber [91] pour palier au problème de "dépendance à long terme". Gardant la même structure en boucle, ils intègrent dans chaque noeud, un ensemble d'opérations supplémentaires qui les différencient des RNNs classiques (voir Fig 2.5). Avec cette structure, les LSTMs ont l'habilité de contrôler le flux d'information à travers notamment les "Gates" qui sont composées d'une couche de sigmoïd et d'un opérateur de multiplication. Autrement dit, les "Gates" permettent de contrôler l'information qu'il faut laisser passer à travers le réseau. Les autres opérations permettent de contrôler les mises à jour

à effectuer et la sortie à produire. Il s'agit là d'une version classique de LSTM qui est souvent modifiée pour résoudre des tâches spécifiques. On peut retrouver des combinaisons de LSTM avec des auto-encodeurs pour faire de la "recontruction-prédiction" pour des tâches de reconnaissance d'actions.

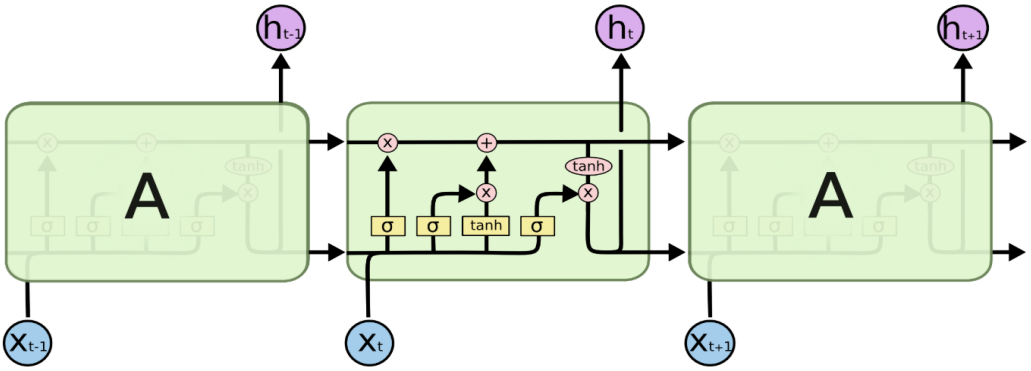


FIGURE 2.5 – Vue détaillée d'un LSTM

2.4.2.2 Le Modèle "LSTM Convolutionnel"

Une variante des LSTMs, utilisée pour la détection d'événements anormaux, est le "convolutional LSTM" (ConvLSTM). Il s'agit d'une composition de LSTM basée sur le modèle des auto-encodeurs avec des couches convolutives, pour modéliser les relations spatio-temporelles entre des images successives. Dans son architecture classique, les couches convolutives successives de LSTM sont utilisées pour encoder les données d'entrée. Les sorties de ces couches servent par la suite à la prédiction de la prochaine image ou région d'image. En utilisant les ConvLSTM comme module dans une architecture globale, Medel et al. [92] ont fait de la détection d'événements anormaux. Dans leur architecture, on peut distinguer deux branches de couches ConvLSTM : une servant à la reconstruction et une autre à la prédiction. L'hypothèse est que le modèle appris à partir du jeu d'apprentissage, serait capable de bien reconstruire et de bien prédire les images contenant des événements normaux. En se basant sur un score de régularité, il leur a été possible d'identifier les images mal reconstruites et mal prédites. Il s'agit des images contenant des événements anormaux.

2.4.3 Les Modèles Génératifs

Il existe d'autres catégories d'approches basées sur les réseaux convolutifs. On peut citer les modèles "génératifs profonds" qui estiment la distribution postérieure conditionnelle de jeu de données dans le but d'entraîner des modèles, capables de générer des données similaires à celles de la base d'apprentissage. Ils apprennent automatiquement des caractéristiques essentielles dans les jeux de données quelque soit la nature et la dimension de ces données. Ils sont de plus en plus utilisés dans des applications variées telles que la prédiction d'actions futures, la traduction de texte, la détection d'anomalies dans le domaine médical, etc. L'un des modèles populaires de cette catégorie est le GAN (Generative Adversarial Networks) qui consiste en un générateur d'images

G et un discriminateur D (souvent un classifieur binaire) qui assigne une probabilité à la donnée générée. Le discriminateur a pour rôle de distinguer une image réelle d'une image fausse (image dont le contenu est de l'aléatoire) tandis que le générateur a pour rôle de générer des images proches des images réelles pour faire croire au discriminateur que ces images sont réelles et non fausses. A la fin de l'apprentissage, une fois les bons paramètres trouvés, le GAN est capable de générer des images dont il sera difficile de dire qu'elles ont été générées. Comparer à d'autres modèles génératifs comme les VAEs (Variational AutoEncoders) ou les modèles auto-régressifs, les GANs génèrent des images plus nettes mais sont plus difficiles à optimiser. L'optimisation se fait à travers le jeu de minmax à deux joueurs suivant la fonction $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))]$$

Des travaux se sont penchés sur l'utilisation des GANs dans la détection d'anomalies. Schlegl et al. [93] ont récemment utilisé le GAN pour la détection d'anomalies dans des images médicales. Le score sur lequel se base la détection, est la probabilité postérieure que l'image de test soit générée par le modèle appris à partir des images de la base d'apprentissage. Ainsi, dans un premier temps, Schlgel et al ont entraîné un modèle GAN à partir d'images médicales sans anomalies. Dans un second temps, pour réellement détecter les anomalies, ils proposent de chercher pour une image de test, le point z dans l'espace latent qui correspond à une image $G(z)$ générée par le GAN et qui serait visuellement similaire à l'image de test x. Pour trouver le meilleur point, ils définissent une fonction de perte composée de deux fonctions de perte : une fonction de perte résiduelle et une fonction de perte de discrimination. La fonction de perte résiduelle est mesurée entre l'image de test et l'image générée et permet de faire converger l'image générée vers l'image de test et la fonction de perte de discrimination permet de s'assurer que l'image générée appartient à l'ensemble d'apprentissage. Cette fonction de perte globale est utilisée dans un processus itératif avec la méthode "back-propagation" afin de trouver le point le plus proche dans l'espace latent. Le score d'anormalité est calculé à partir de ces fonctions de perte suivant l'expression :

$$A(\times) = (1 - \lambda).R(\times) + \lambda.D(\times)$$

avec $R(\times)$ le score résiduel et $D(\times)$ le score de discrimination. Un score élevé signifie que l'image de test n'a pas participé à l'apprentissage du modèle GAN et donc correspond à une image contenant une anomalie.

2.5 Conclusion

Ce chapitre a été consacré à la revue générale de la littérature sur la détection d'événements rares ou anormaux dans une vidéo. L'intérêt pour cette thématique est grand et ne date pas d'hier comme nous avons pu le constater à travers le résumé des travaux présentés. Les premiers travaux se sont positionnés sur la détection globale des événements rares dans une scène sans se préoccuper de leur localisation précise. Le défi était de détecter le début et la fin des événements rares dans un flux vidéo. L'application favorite de ces algorithmes est la détection d'événements dans une scène de densité élevée de personnes. Les premiers travaux ont commencé dans les domaines de la physique pour modéliser les comportements des foules à des fins de simulations. Très vite, avec le développement des algorithmes de "machine learning" et de la vision par ordinateurs, de nouveaux travaux inspirés des méthodes de la physique ont vu le jour.

Les nouveaux défis se trouvant plus au niveau de la localisation exacte de l'événement suspect dans la scène, plusieurs travaux ont récemment porté sur cette thématique. Nous avons présenté plusieurs approches qui sont soit inspirées de la physique ou qui se basent directement sur des algorithmes avancés, utilisés pour d'autres tâches en vision par ordinateur et en "machine learning".

Toutes ces méthodes présentent généralement des frameworks similaires et se différencient sur deux aspects cruciaux : les informations extraites et l'algorithme de modélisation utilisé. La qualité des informations est primodiale pour une bonne détection. Elle dépend du type de l'information à extraire et de comment l'extraire. Sur ce plan, nous avons donc noté de nombreux descripteurs dont le plus utilisé reste le "flot optique". Chaque méthode propose un schéma d'extraction de ces informations qui sont soit uniquement spatiales ou soit spatio-temporelles. La modélisation fait appel la plupart du temps à des algorithmes de classification de données tels que les SVM, les modèles de Markov, les classifieurs bayésiens et autres.

Une toute nouvelle ère s'ouvre dans le domaine de "machine learning" avec l'apparition des approches basées sur les réseaux de neurones convolutifs. Ces approches présentent, dans plusieurs domaines, une forte attraction due aux résultats impressionnants qu'elles obtiennent. Des chercheurs ont très récemment, commencé à se pencher sur l'utilisation de ces dernières pour les tâches de détection et la localisation des événements anormaux. Nous avons, dans ce chapitre, présenté une gamme variée d'approches basées sur les réseaux de neurones convolutifs ainsi que leur utilisation pour la détection des événements ou d'anormalités dans des images médicales.

Ce chapitre n'a pas pour but de présenter toutes les méthodes et approches existantes, mais de présenter de façon sommaire la grande variété d'approches qui existent. Ainsi, il reste dans chaque catégorie, de nombreuses autres approches et d'inombrables travaux. Ceux présentés ici, ont servi de point de départ pour les contributions faites durant cette thèse.

Chapitre 3

Analyse globale de scène pour la détection d'événements rares : cas de panique de foule

> Vision is the process of discovering from images what is present and where it is.
>
> David Marr

Sommaire

3.1 Introduction

Comme nous l'avons montré dans le chapitre 2, il existe plusieurs approches abordant la détection de différents types d'événements dans la littérature. Nous pouvons distinguer les événements globaux et ceux locaux. Le niveau d'extraction et de modélisation des informations de la scène se définit en fonction donc de l'application visée. Dans le cas des événements globaux, une localisation n'est pas nécessaire mais ce qui importe est la détection du début et de la fin de leur déroulement.

Dans ce chapitre, nous aborderons le problème de la détection d'événements globaux entraînant une dynamique globale de la scène. Notre approche, baptisée BGMMAI pour "Bayesian Generative Model based on Motion and Appearance Information", peut être classée dans la catégorie des approches "sac-de-mots". Nous proposons d'utiliser une modélisation bayésienne des évènements à l'aide de l'algorithme LDA en utilisant comme caractéristiques les informations de mouvement et d'apparence simultanément.

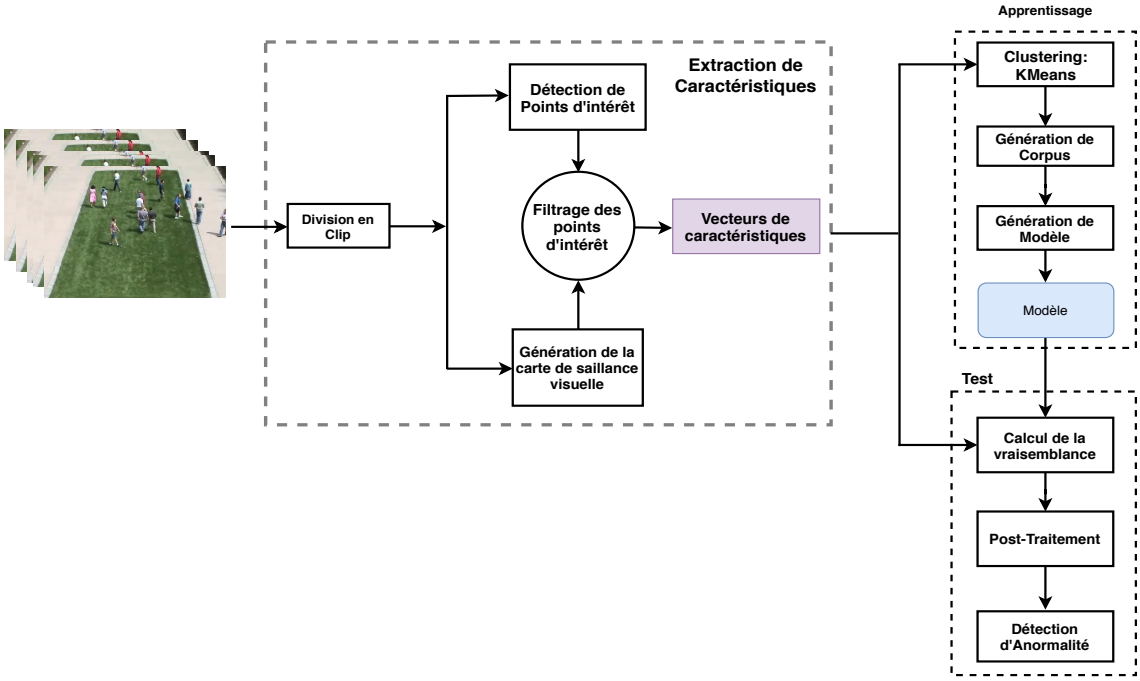


FIGURE 3.1 – Méthodologie de l'approche BGMMAI

La figure 3.1 donne une vue globale de la méthode que nous proposons pour la détection des scènes de panique de foule. Le traitement des événements est fait de façon temporaire. Dans un premier temps, le flux vidéo entrant est divisé en des séquences d'images appelées "clip". Le traitement se fait alors "clip" par "clip". Dans chaque clip, on procède à la mise en évidence des entitées pertinentes qui y sont présentes, grâce à des algorithmes de saillance visuelle. A cette étape, chaque pixel de l'image a une probabilité de saillance. Parallèlement à la mise en évidence des régions saillantes, un ensemble de points d'intérêt spatio-temporel (STIP) est extrait dans la scène. Ces points d'intérêt peuvent se localiser sur toutes entités (pertinentes visuellement ou non) présentes dans la scène. En partant du principe que généralement, les événements seront

générés par des entitées pertinentes, un filtrage des points d'intérêt est alors appliqué afin de ne retenir que ceux détectés dans les régions saillantes. La description de la scène peut alors commencer par une extraction successive d'informations de mouvements et d'apparence aux alentours des points d'intérêt retenus après filtrage. Plusieurs descripteurs existent dans la littérature dont les plus récents et pertinents ont été présentés dans le chapitre 1. Nous utiliserons dans notre méthode deux descripteurs : l'histogramme des orientations du flot optique (HOOF) pour le mouvement et l'histogramme des orientations de phase (CHOP) pour l'apparence. A partir de l'ensemble des informations issues de la description de la scène, nous procédons à la construction du modèle final d'événements grâce à l'algorithme "Latent Dirichlet Allocation" (LDA) largement utilisé avec succès dans le domaine de fouille documentaire mais également dans la détection d'événements. L'utilisation de cet algorithme requiert des entrées sous le format "mots visuel-document" encore connu sous le vocable "bag of words". Ainsi, suivront une succession d'algorithmes pour transformer les informations précédemment extraites en "mots-visuels" puis en document.

3.2 Filtrage de points d'intérêt basé sur la saillance visuelle

3.2.1 Les points d'intérêt

La notion de points d'intérêt a été introduite pour la première fois en traitement d'image par Moravec [94] dans les années 1977. Il les décrit comme étant des points dans une image où l'on peut observer une forte variation de l'intensité lumineuse dans au moins deux directions, contrairement aux contours qui sont un ensemble de points avec des variations unidirectionnelles. Les points d'intérêt sont abondamment utilisés en vision par ordinateur pour diverses applications telles que la reconnaisance d'objets, le tracking, etc. Une riche littérature existe sur le sujet [94, 95, 96, 97].

Soit I une image quelconque et $p = (x, y)^T$ un point donné dans l'image. Le calcul des variations locales de I en p associées a un déplacement $\Delta p = (\Delta x, \Delta y)$ est donné par la fonction d'autocorrélation suivante :

$$\chi(p) = \sum_{p \in W} (I(p) - I(p + \Delta p))^2 \tag{3.1}$$

avec W une fenêtre centrée au point p.

Le terme $I(p + \Delta p)$ peut être détaillé par une approximation du premier ordre comme suit :

$$I(p + \Delta p) \simeq I(p) + \left(\frac{\partial I}{\partial x}(p) \quad \frac{\partial I}{\partial y}(p) \right) . \Delta p$$

En intégrant cette approximation dans l'équation 3.1, on obtient :

$$\begin{aligned} \chi(p) &= \sum_{p \in W} \left[\left(\frac{\partial I}{\partial x}(p) \quad \frac{\partial I}{\partial y}(p) \right) . \Delta p \right]^2 \\ &= \Delta p^T . M(p) . \Delta p \end{aligned}$$

avec $M(p)$ la matrice d'autocorrélation représentant les variations locales de I en p :

$$M(p) = \begin{pmatrix} \sum_{(x_k,y_k)\in W} (\frac{\partial I}{\partial x}(x_k,y_k))^2 & \sum_{(x_k,y_k)\in W} \frac{\partial I}{\partial x}(x_k,y_k).\frac{\partial I}{\partial y}(x_k,y_k) \\ \sum_{(x_k,y_k)\in W} \frac{\partial I}{\partial x}(x_k,y_k).\frac{\partial I}{\partial y}(x_k,y_k) & \sum_{(x_k,y_k)\in W} (\frac{\partial I}{\partial y}(x_k,y_k))^2 \end{pmatrix}$$

$\chi(p)$ étant la fonction des variations locales, avec la matrice d'autocorrélation $M(p)$ décrivant sa forme à l'origine (explicitement, les termes quadratiques dans l'expansion de Taylor) [98]. En se basant sur les valeurs propres de la matrice $M(p)$, quelques propriétés intéressantes ont été mises en évidence. Soit α et β deux valeurs propres de $M(p)$. Elles sont proportionnelles aux courbures principales de $\chi(p)$. Les trois propriétés mises en évidences sont les suivantes :

- si les deux courbures sont petites, alors le point considéré se trouve dans une région homogène,

- si une courbure est plus grande que l'autre, alors le point se trouve sur un contour,

- si finalement, les deux courbures sont grandes (donc, une variation de l'intensité dans les deux directions), le point correspond donc à la définition de Moravec.

Ainsi, en observant la catégorisation dans l'espace (α, β) (Fig 3.2), on peut dire qu'un point d'intérêt doit avoir de grande valeur pour α et β.

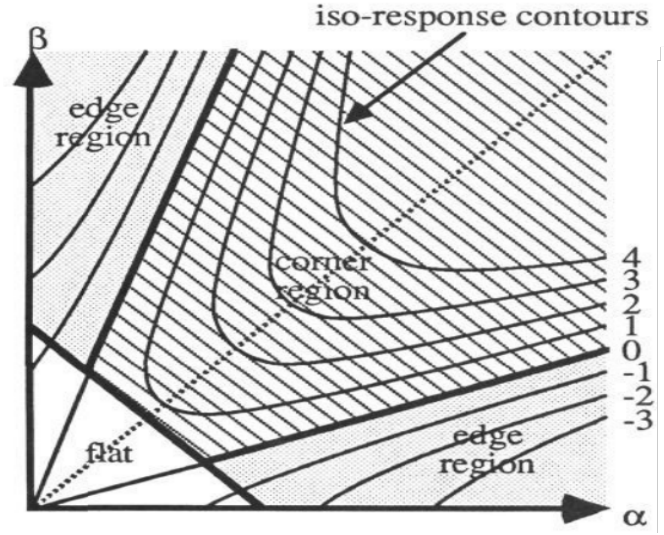


FIGURE 3.2 – Classification des points dans l'espace des vecteurs propres [98]

Le plus connu et utilisé parmi les détecteurs de points d'intérêt dans la littérature est celui de Harris-Stephens [98]. Dans leurs travaux, Ils établissent une relation de courbure R afin de s'affranchir de la détermination des valeurs propres de M :

$$R = Det(M) - k.Trace^2(M) \tag{3.2}$$

$M(p)$ peut être vu sous la forme :

$$M = \begin{pmatrix} A & C \\ C & B \end{pmatrix}$$

Ainsi, $Det(M) = \alpha\beta = AB - C^2$ et $Trace(M) = \alpha + \beta = A + B$. La valeur de k est une constante et est obtenue de façon empirique. Harris et Stephens déduisent que pour $R >> 0$, le point considéré est un point d'intérêt.

Les points d'intérêt ainsi extraits dans la scène, se localiseront autour de toute entité (personne) ou objets qui y sont présents. Ces entités ne participant pas forcément à la génération d'évènements, nous proposons d'utiliser l'information de saillance visuelle dans la sélection de points d'intérêt pertinents. Nous avons utilisé le détecteur de Harris-Stephens dans les travaux sur la méthode BGMMAI.

3.2.2 La saillance visuelle

La saillance, en anglais "saliency", est selon le dictionnaire en ligne l'internaute [1], une chose qui avance, qui dépasse au point d'être vue, qui peut être remarquée par contraste par rapport à un alignement général. Ou encore une chose qui a une forte possibilité d'être remarquée. Partant de cette définition, la "saillance visuelle" peut-être définie comme étant des régions d'intérêt facilement remarquables par la vision humaine.

En "vision par ordinateur", les scientifiques proposent des algorithmes capables de mettre en évidence automatiquement dans une image, toutes entités présentant de forte probabilité d'être aperçues par la vision humaine. Cette tâche est réalisée ultra rapidement par le cerveau humain [99] mais demeure un vrai défi pour les scientifiques, tant sur le plan de la précision de la détection que sur le temps de calcul. De nombreux travaux [100, 101, 102, 103] ont été menés pour arriver à élaborer des algorithmes efficaces et rapides. Ces méthodes, à contrario de celles de segmentation, n'ont pas pour objectif de faire une classification multi-classes des pixels de l'image, mais bien une classification binaire avec une probabilité de saillance attribuée à chaque pixel.

Dans une image, les zones qui se ressemblent peuvent être perçues comme une redondance d'information contenue dans celle-ci et qui ne sont rien d'autre que l'arrière-plan. Une zone contenant un objet saillant est donc une nouvelle information ajoutée à l'information redondante (insertion d'un objet sur l'arrière-plan). Soit $H()$ l'opérateur qui exprime l'information contenue dans un signal. L'expression de l'information contenue dans une image est :

$$H(Image) = H(innovation) + H(redondance)$$

La mise en évidence des régions saillantes reviendrait à supprimer l'information redondante $H(redondance)$ pour ne garder que l'information nouvelle $H(innovation)$.

Hou et al. [102], ont présenté un algorithme de calcul de la saillance basé sur l'analyse des spectres de Fourier de l'image. Ce travail a été motivé par les découvertes faites par Oppenheim et Lim [103] lors de travaux sur l'analyse de l'importance de la phase dans les signaux. Dans leurs travaux, ils ont montré que la reconstruction d'un signal à partir uniquement de la phase de sa transformée de Fourier est plus proche de l'image originale qu'une copie issue d'une reconstruction basée uniquement sur l'amplitude. Cela montre qu'une partie importante de l'information du signal est contenue dans la phase de la transformée de Fourier. L'idée proposée par Hou et al. est de construire une nouvelle image à partir de la phase et de l'amplitude du spectre résiduel de la transformée de Fourier de l'image originale. Le spectre résiduel équivaut au reste du spectre

1. `www.linternaute.fr/dictionnaire/fr/definition/saillance/#definition`, consulté le 07/12/2018

après y avoir soustrait le spectre moyen. En d'autres termes, $\mathcal{R}(f)$ désigne les singularités statistiques propres à l'image d'entrée qui contiennent plus les hautes fréquences que les basses. Cette opération est effectuée dans l'espace log des spectres (largement utilisées en analyse statistique d'images).

Soit $\mathcal{I}(x, y)$ une image, contenant des régions d'intérêt à détecter. L'algorithme proposé par Hou et al.[102] est la suivante :

Algorithme 1 : Calcul de la saillance basée sur l'analyse des spectres de Fourier

Entrées : $\mathcal{I}$: Image originale

Output : $S(\mathcal{I})$: Carte de saillance de l'image d'entrée

1 Calculer le spectre $\mathcal{L}(f) = log(\Re(\mathcal{F}[\mathcal{I}]))$ de la transformée de Fourier de l'image;

2 Calculer la phase $\mathcal{P}(f) = \Im(\mathcal{F}[\mathcal{I}])$ de la transformée de Fourier de l'image;

3 Calculer l'amplitude du spectre moyenné de Fourier $\mathcal{A}(f)$ de l'image;

4 Calculer le spectre résiduel $\mathcal{R}(f) = \mathcal{L}(f) - \mathcal{A}(f)$ de l'image;

5 Construire la carte de saillance : $S(\mathcal{I}) = g * \mathcal{F}^{-1}[exp(\mathcal{R}(f) + \mathcal{P}(f))]^2$.

$\mathcal{A}(f)$ peut être approximé par $\mathcal{A}(f) = h_n * \mathcal{L}(f)$ où h_n est un filtre moyen de taille n x n défini par : $h_n = \frac{1}{n^2}\begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ . & . & . & . \\ . & . & . & . \\ 1 & 1 & ... & 1 \end{pmatrix}$. Ce qui donne : $\mathcal{R}(f) = \mathcal{L}(f) - h_n * \mathcal{L}(f)$.

La carte de saillance $S(\mathcal{I})$ est alors la construction d'une image à partir de la transformée inverse du spectre résiduel et de la phase du spectre de l'image originale, avec un lissage gaussien pour un meilleur effet de visualisation. On peut observer avec la figure 3.3 quelques résultats de la méthode tirée de [102].

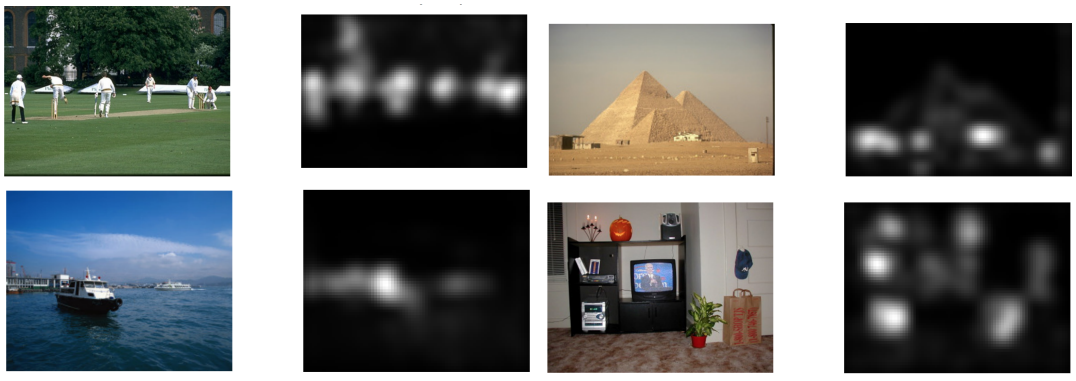


FIGURE 3.3 – Example de calcul de la saillance [102]

3.2.3 Filtrage des points d'intérêt

La modélisation des événements nécessite une bonne sélection des caractéristiques de la scène. La méthode BGMMAI étant basée sur une approche "sac de mots", l'utilisation des points

d'intérêt favorise la focalisation sur des objets présents dans la scène pour l'extraction localisée des caractéristques. Les points d'intérêt peuvent être plusieurs dans la même zone, mais également peuvent être localisés sur des objets non pertinents visuellement. Cela peut avoir deux effets non désirables sur le modèle d'événements à apprendre :

- la redondance d'informations qui n'apporte rien de plus au modèle mais peut alourdir le temps de calcul,

- l'introduction d'une erreur dans le modèle en prenant en compte des entités ne rentrant pas en compte dans la détection d'événements. Apprendre l'arrière-plan dans la création du modèle, par exemple, impliquera à coup sûr des erreurs et impactera les performances de la méthode.

Nous proposons dans cette méthode, une approche de selection simple mais efficace des points d'intérêt. Cette approche se base sur deux hypothèses à savoir :

- les entités générant des événements pouvant être prises en compte dans la modélisation, sont visuellement saillants,

- un ensemble de points d'intérêt se localise sur des entités qui peuvent être pertinentes ou non dans la génération d'événements.

Notre approche est décrite par l'algorithme ci-dessous :

Algorithme 2 : Selection de points d'intérêt pertinents

Entrées :

- $\mathcal{L}(pt_i)$: Liste de points d'intérêt intiaux

- $S(p)$: Carte de saillance des images du clip

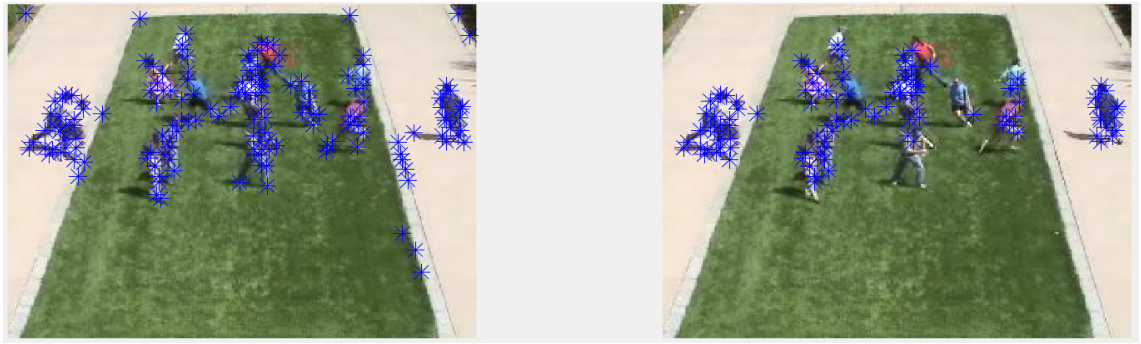
Output : $\mathcal{L}(pt_o)$: Liste de points d'intérêt filtrés

```
pour p_i ∈ L(pt_i) faire
    Prendre une cuboïd f_i centrée sur p_i;
    Calculer la moyenne m de la saillance de f_i;
    Calculer la variance v de la saillance de f_i;
    si (m > th_moy) et (v > th_var) alors
        Ajouter p_i dans L(pt_o);
        Supprimer tous les p_i ⊂ f_i de la liste L(pt_i);
    sinon
        Supprimer p_i de L(pt_i);
```

Trois paramètres interviennent dans l'algorithme. Il s'agit de la taille de la fenêtre spatio-temporelle glissante (cuboïd), le seuil sur la moyenne et le seuil sur la variance. La taille temporelle du cuboïd est égale à la taille du clip et la taille spatiale est la même que celle utilisée pour l'extraction de caractéristiques autour des points. Quant aux seuils sur la moyenne et la variance, ils sont déterminés empiriquement. L'objectif de l'approche est de choisir dans une petite fenêtre, contenant des points saillants, un seul point d'intérêt dont les caractéristiques seront sensiblement égales à celles des autres points qui s'y trouvent. La moyenne, permet de s'assurer que le point est réellement saillant tandis que la variance, permet de s'assurer de la saillance

temporelle du point, afin d'éliminer les fausses alertes. La figure 3.4 montre un exemple de résultat de filtrage de point d'intérêt sur dans une scène. On peut voir sur les images ci-dessous, que des points d'intérêt non pertinents se trouvant sur les bords et coins des pelouses sont filtrés.

a



b

FIGURE 3.4 – Exemple de résultat de filtrage de points d'intérêt :
a : A gauche l'image d'entrée et à droite la carte de saillance
b : A gauche image avant filtrage des points d'intérêt et à droitre après filtrage

3.3 Extraction de caractéristiques par la description des points d'intérêt

Autour de chaque point d'intérêt retenu à l'issue du filtrage, dans une fenêtre de taille donnée, l'extraction de caractéristiques locales vise à extraire des informations de différents types qui décrivent un phénomène, une grandeur, à l'échelle locale. Dans le cadre des travaux de cette thèse, la nature des événements imposent que nous nous focalisions sur des caractéristiques liées au mouvement et éventuellement à la forme des différentes entités de la scène. Différents descripteurs de scène de la littérature ont été présentés (voir chapitre 1). Parmi les descripteurs de la littérature, la plupart ont déjà été utilisés dans des travaux de détection d'événements avec des performances variées.

Pour cette approche, nous avons fait appel à deux descripteurs différents utilisés séparemment ou de façon combinée pour montrer leur pertinence pour le type d'événement ciblé. Pour la caractérisation du mouvement, le flot optique est utilisé. Pour ce qui est de l'apparence, les descripteurs étudiés dans le chapitre 1 ont été utilisés. Le choix de ces descripteurs est basé sur leur robustesse dans certaines situations auxquelles on pourra être amené à faire face.

3.3.1 Description du mouvement dans la scène : Histogramme des Orientations du Flot Optique

L'histogramme des orientations du flot optique s'inspire de différents descripteurs sous forme d'histogramme qui ont connus un succès dans le domaine de la "vision par ordinateur". Il a été introduit en vision par Chaudhry et al. [104] et permet d'avoir une vue globale de la distribution des informations de vitesse et de direction des pixels dans une image donnée. Le calcul du flot optique repose sur l'hypothèse fondamentale selon laquelle l'intensité lumineuse se conserve entre deux images successives. Cette hypothèse vient du faite qu'on ne peut estimer le déplacement d'un objet dont les pixels n'ont pas la même intensité entre deux images, car estimer le déplacement d'un objet, revient à estimer le déplacement des pixels. La formulation de cette hypothèse est la suivante :

$$I(x, y, t) - I(x + dx, y + dy, t + dt) = 0 \tag{3.3}$$

avec (dx, dy) la distance du déplacement des pixels d'une image à une autre. Sous la forme différentielle en utilisant l'approximation de la série de Taylor, l'équation 3.3 peut être exprimée comme suit :

$$\frac{dI(x(t), y(t), t)}{dt} = \frac{\partial I}{\partial x}.\frac{dx}{dt} + \frac{\partial I}{\partial y}.\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \tag{3.4}$$

soit

$$I_x u + I_y v + I_t = 0$$

où $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, $I_t = \frac{\partial I}{\partial t}$, $u = \frac{dx}{dt}$ et $v = \frac{dy}{dt}$.
Enfin, sous une forme compacte, on a :

$$(\nabla I)^T \omega + I_t = 0 \tag{3.5}$$

avec $\nabla I = [I_x, I_y]^T$ le gradient spatial, I_t, la dérivée temporelle et $\omega = [u, v]^T$, le flot optique recherché. L'estimation de ω est impossible à partir de la seule hypothèse fondamentale car il y a deux inconnus à estimer. L'ajout d'une contrainte supplémentaire est nécessaire pour résoudre le problème. Ainsi, différentes méthodes émettent des contraintes visant à minimiser une fonctionnelle basée sur l'équation 3.5.

Dans les année 1981, deux méthodes ont vu le jour et sont abondamment utilisées dans la littérature : l'algorithme de Horn-Schunck [105] et celui de Lucas-Kanade [106]. Horn et Schunk, dans leur méthode, proposent une minimisation sur l'ensemble de l'image de la fonctionnelle ci-après, contenant une contrainte de régularisation sur le gradient, en utilisant l'opérateur de Laplace :

$$E = \int\int ((\nabla I)^T \omega + I_t)^2 + \alpha^2((\nabla v_x)^2 + (\nabla v_y)^2) dx dy \tag{3.6}$$

Cette méthode donne des résultats de flot lisse et est très sensible aux bruits à cause du terme de lissage supplémentaire sur l'ensemble de l'image. Elle élimine également les petits mouvements se produisant dans la scène.

A contrario, Lucas et Kanade ont proposé une méthode locale, travaillant sur une petite fenêtre spatiale Ω de taille $n \times n$ autour d'un pixel. Ils supposent que tous les pixels dans la

fenêtre ont des mouvements similaires. A partir cette contrainte supplémentaire, ils ont cherché à minimiser la fonctionnelle ci-après, en utilisant une fenêtre pondérée W avec le critère des moindres carrés.

$$E = \sum_{\Omega} W^2 [\nabla I.\omega + I_t]^2 \tag{3.7}$$

Il est intéressant de noter que l'expression de Euler-Lagrange de l'équation 3.7 est de la forme $v = H^{-1}.b$:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum\limits_{\Omega}(I_x)^2 & \sum\limits_{\Omega} I_x.I_y \\ \sum\limits_{\Omega} I_x.I_y & \sum\limits_{\Omega}(I_y)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum\limits_{\Omega} I_x I_t \\ -\sum\limits_{\Omega} I_y I_t \end{bmatrix}$$

On peut ainsi noter la ressemblance entre la matrice H et celle de corrélation des points de Harris. Cela permet de déduire qu'il serait bon de suivre uniquement les coins pour l'estimation du flot optique dans une vidéo. Cette méthode est moins sensible aux bruits à cause de son traitement local.

Dans sa version originale, Chaudhry et al. [104] ont utilisé le descripteur HOOF (Histogram of Oriented Optical Flow) pour la reconnaissance d'action. Pour une telle application, la prise en compte de toutes les directions n'est pas nécessaire pour distinguer les mouvements. Ainsi, les mouvements dans les directions gauches et droites sont regroupés ensemble. En effet, la construction de l'histogramme consiste à diviser l'espace des orientations en B intervalles (bins) réguliers. Chaque pixel de l'image vote pour un "bin" donné en fonction de son angle principal par rapport à l'axe horizontal (voir Fig 3.5). Soit $\omega = [u, v]^T$ le vecteur représentant le flot optique, $\theta = tan^{-1}(\frac{v}{u})$ sa direction et b le "bin" pour lequel il devra voter. Ces trois données (b, B, θ) sont liées par la relation suivante :

$$\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{B}$$

Le vote de chaque flot est pondéré par son amplitude ($\sqrt{u^2+v^2}$) afin de prendre en compte cette information captiale, la vitesse, pour distinguer par exemple une marche d'une course. Une normalisation de l'histogramme permet de le rendre invariant à l'échelle.

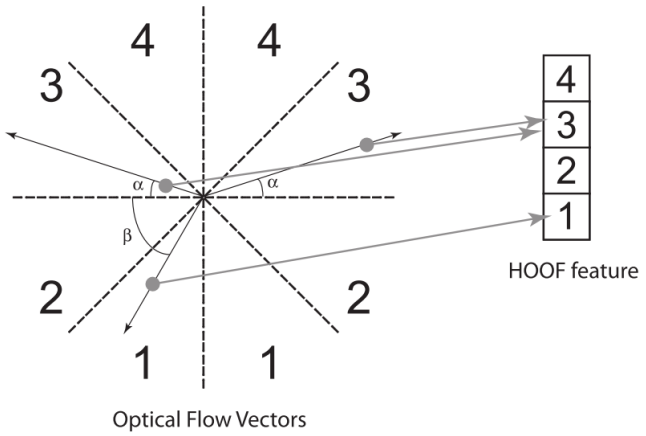


FIGURE 3.5 – Découpage de l'espace des orientations dans la version originale [104]

Pour la détection d'événements, il est indispensable de prendre en compte toutes les directions. Ainsi, nous proposons une modification du vote des bins pour l'adapter à notre situation (voir

Fig 3.6). De plus, l'invariance à l'échelle est discutable. En fonction de l'application, elle peut ne pas être nécessaire. Prenons l'exemple de la surveillance d'une zone donnée. Il peut arriver que deux mouvements identiques, dans des zones différentes de la scène ne soient pas considérés de la même manière. Dans l'une, il peut s'agit d'un événement normal et dans l'autre, d'un événement rare. Cette contrainte nous amène à laisser l'histogramme sans normalisation.

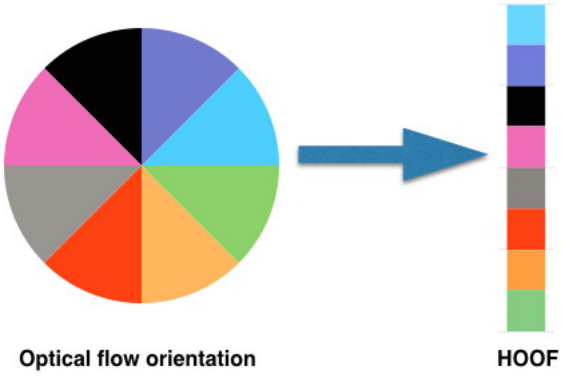


FIGURE 3.6 – Découpage de l'espace des orientations dans la version modifiée

Pour extraire le descripteur HOOF, nous considérons l'ensemble des points d'intérêts retenus après filtrage. Autour de chaque point, dans une fenêtre de dimension donnée, nous calculons HOOF pour chaque image du clip considéré. Ainsi, pour un point p, nous aurons un ensemble ni (nombre d'image du clip) de vecteurs : $F_p = \{f_1, f_2, ..., f_{ni}\}$. Différents schémas peuvent être envisagés pour combiner ces vecteurs. On peut envisager une concaténation, une moyenne par "bin", une somme ou une somme pondérée en attribuant des poids à chaque image du clip. Les différentes expériences menées montrent qu'il n'y a pas une grande différence entre ces différents schéma.

3.3.2 Description de l'apparence dans la scène

En plus de l'information de mouvement, celle d'apparence des entités présentes, peut être exploitée. Comme nous l'avons mentionné plus haut, le choix des informations à exploiter est hautement dépendant de l'application visée. Pour une application basée uniquement sur les mouvements, la prise en compte d'une telle information s'avère presque inutile. Notre approche se veut être générale pour inclure le maximum d'événements. Dans le chapitre 1, nous avons présenté des descripteurs qui nous ont servi à détecter des personnes et des objets dans des scènes avec comme contrainte une variation de la luminosité. De cette étude, nous avons montré l'efficacité du descripteur CHOP, basé sur la congruence des phases des images dans le domaines de Fourier. Ce descripteur a montré sa supériorité sur le descripteur HOG, abondamment utilisé, dans un contexte nocturne. Il est à noter qu'à ce jour, aucun travail publié n'utilise le descripteur CHOP dans le contexte de la détection d'événements. Une étude comparative de l'apport réel de chacun des descripteurs CHOP et HOG sera présentée dans la partie évaluation de ce chapitre.

3.4 Modélisation d'événements

La modélisation des événements revient à construire un modèle qui englobe toutes les informations issues du jeu d'apprentissage afin de servir à la classification des jeux de test. Un bon algorithme d'apprentissage tend à améliorer la représentation des données d'apprentissage en utilisant un certain nombre de paramètres dont les valeurs sont à déterminer. Il doit être robuste au bruit et rapide en temps de calcul. Plusieurs algorithmes sont utilisés dans la littérature pour répondre à ce genre de problème : du simple matching, par calcul de distance géométrique entre vecteurs de caractéristiques, aux plus élaborés des algorithmes d'apprentissage. Le problème de classification que nous traitons ici est un problème de classification binaire avec une seule classe lors de l'apprentisage ("one class"). Autrement dit, l'algorithme de classification construira un modèle qui définira un contour autour du jeu d'apprentissage afin d'éjecter toutes données qui s'y éloignent à partir d'un seuil de distance donné. Nous avons décidé pour cette méthode, d'utiliser l'algorithme d'Allocation latente de Dirichlet (LDA : Latent Dirichlet Allocation) qui est un algorithme non supervisé de découverte de thème dans un ensemble de documents traitant de thématiques variées. Cette découverte non supervisée des différentes thématiques traitées par les documents du corpus peut être adaptée dans le cadre de la modélisation des événements dans une image. L'hypothèse fondamentale est que les événements, même étant jugé normaux, ne sont pas de même nature : on parlera de différences intra-classes. La découverte du thème peut être assimiliée à la découverte de ces groupes d'événements, tous jugés normaux. L'utilisation de cette méthode en vision par ordinateur, ne date pas d'aujourd'hui. Hu et al. [107] ont montré que le LDA peut être aussi utilisé dans le domaine de l'image et de la musique. Il s'agit le plus souvent de faire de la classification multi-classe, où chaque classe correspond à un thème. Ainsi, la classification d'objets regroupés en catégories, peut être réalisée en utilisant cet algorithme. La décision d'attribution d'un vecteur de caractéristiques de test à un thème se fait en regardant les probabilités obtenues par thème. Pour la classification "one-class", c'est la vraissemblance du vecteur de caractéristiques à classer qui est utilisée. La vraissemblance exprime, en fonction des probabilités des thèmes apparus dans le document test, de combien le document ressemble au corpus appris.

L'utilisation du LDA, impose une représentation en "sac-de-mots" ("bag-of-word" en anglais) des clips vidéos dont le contenu doit être classé en normal ou rare. Un clip est une séquence de n images regroupées ensemble. La taille d'un clip est égale au nombre d'images qui appartient à ce dernier. Pour une vidéo donnée, les clips peuvent être extraits en utilisant une fenêtre glissante avec ou sans chevauchement. La figure 3.7 illustre le découpage d'une vidéo en des clips de taille 3. La représentation en "sac-de-mots" a été utilisée pour plusieurs tâches dans la littérature dont celles de détection et de classification d'objets. Le processus de représentation d'une image en "sac-de-mots" comporte généralement trois phases : l'extraction des caractéristiques, le clustering et l'identification des centroïds et la représentation sous forme d'histogramme à travers un processus de vote. Dans cette section, nous détaillerons les étapes de la modélisation des événements, en partant de la représentation en "sac-de-mot" jusqu'au calcul de la vraissemblance des

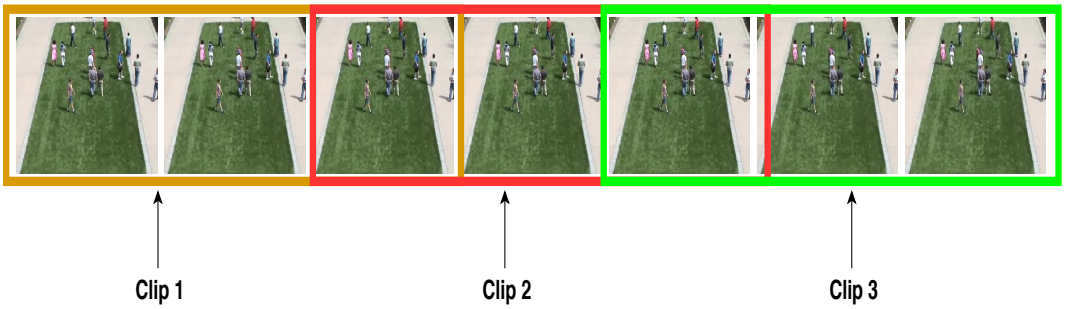


FIGURE 3.7 – Illustration de découpage des clips

clips qui servira de score de décision pour notre classification.

3.4.1 Représentation en sac-de-mots des caractéristiques

A la fin de l'étape de détection et de description des points d'intérêt spatio-temporels, nous disposons pour chaque clip, d'un ensemble de vecteur de descripteurs. A l'étape de clustering, l'ensemble des vecteurs de descripteurs de tous les clips est partitionné en un ensemble de sous-groupe (appelé clusters). Chaque sous-groupe est représenté par un vecteur qui, géométriquement représente le barycentre de l'ensemble des vecteurs qui appartiennent au sous-groupe (on parle de centroïd). Pour identifier les clusters et leur centroïd, on fait appel aux algorithmes de partitonnement de données. Dans notre méthode, nous avons opté pour l'algorithme K-Means (encore appellé K-Moyennes) avec la distance de Bhattacharyya étant donné qu'on utilise des histogrammes.

Le nom "K-Means" attribué à l'algorithme, a été utilisé pour la première fois en 1967 par James MacQueen [108]. L'idée principale de l'algorithme est de minimiser la somme des carrées des distances de chaque points à la moyenne des points appartenant à son cluster. En cherchant à minimiser les distances entre les points et les centroïdes, on obtient à la fin, des clusters composés de points toujours plus proches du centroïd du cluster d'appartenance que des centroïds des autres clusters. Il s'agit d'un algorithme simple et efficace bien que ne garantissant ni l'optimalité, ni un temps de calcul polynominal [109]. Soit un ensemble P de p points répartis dans l'espace et $C = \{C_1, C_2, ..., C_k\}_{(k \leq p)}$ les k sous-ensembles qui représentent les clusters à identifier. L'algorithme 3 présente le déroulement de l'identification des clusters.

L'étape d'initialisation de l'algorithme est très importante car elle influence le temps d'exécution mais également la qualité des clusters obtenus. Des travaux ont été menés sur le choix efficace des points initiaux. Deux méthodes sont souvent utilisées à cet effet. D'une manière simple, Forgy et al. [110] proposent de choisir des points au hasard parmi les points à partitionner. La seconde méthode consiste à assigner aléatoirement un cluster à chaque point et par la suite calculer les centroïds initiaux en se basant sur les clusters de chaque point. Pour limiter le temps de calcul de l'algorithme, qui peut être exponentiel au nombre de points, dans la pratique, il est fixé une limite en terme d'itérations ou un critère de convergence qui mesure l'amélioration du partitionnement entre deux itérations successives. L'image de la figure 3.8 montre un exemple

du déroulement de l'algorithme pour le partionnement de points dans l'espace 2D. Les points peuvent appartenir à des espaces de dimensions quelconque. Dans notre méthode, les points à partitionner étant les vecteurs de descripteurs qui sont des histogrammes, nous avons utilisé la distance de Bhattacharyya comme métrique pour l'assignation des points aux clusters. Le seul paramètre important à définir est le nombre k de clusters à déterminer. Le choix d'une valeur fixe de ce paramètre peut en fonction du domaine d'application être un inconviénient ou un avantage. Dans le cas spécifique des approches "sac-de-mots", cela est un avantage puisqu'il permet d'avoir une taille fixe de dictionnaire. Le nombre de mots dans un dictionnaire pour de telle application est fixé par défaut pour avoir des vecteurs de caractéristiques formés à partir des mêmes mots que ce soit dans la phase d'apprentissage ou de test.

Algorithme 3 : Algorithme K-Means

Entrées :
- P : Ensemble des points
- k : Le nombre de clusters à identifier

Output :
C : Ensemble des clusters

1 Initialisation : Choisir k points représentant les moyennes des partitions $m_1^{(0)}, ..., m_k^{(0)}$;
2 **tant que** *L'assignation de cluster n'est pas stable* **faire**
3 Assigner chaque point au cluster le plus proche;
4 Partitionner l'espace en cluster à partir des moyennes $m_i^{(n)}$. On peut utiliser le partitionnement de Voronoï : $C_i^{(n)} = \{x_j : ||x_j - m_i^{(n)}|| \leq ||x_j - m_{i^*}^{(n)}|| \ \forall \ i^* = 1, .., k\}$;
5 Mettre à jour la moyenne de chaque cluster : $m_i^{(n+1)} = \frac{1}{|C_i^{(n)}|} \sum_{x_j \in C_i^{(t)}} x_j$;
6 Incrémenter n ;

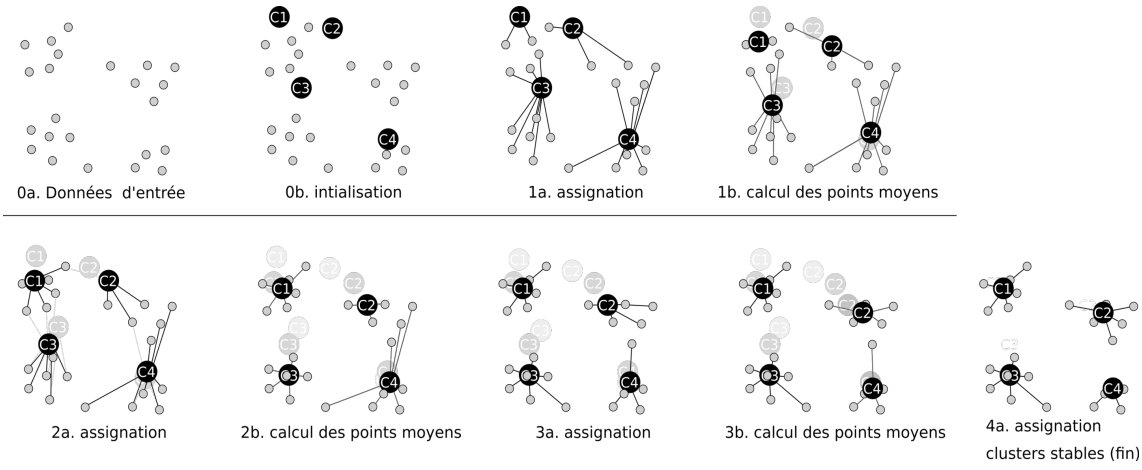


FIGURE 3.8 – Exemple de déroulement de l'algorithme K-Means. source : Par Mquantin — Travail personnel, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=61321400`

A l'issue de l'exécution de l'algorithme 3 sur l'ensemble de nos vecteurs de descripteurs, nous obtenons k centroïds ainsi que le cluster d'appartenance de chaque vecteur. L'étape suivante, consiste à construire pour chaque clip, un histogramme d'occurence d'apparition de chaque cen-

troïd. La construction de l'histogramme revient à calculer pour chaque cluster, le nombre de points d'intérêt lui appartenant :

$$|c_i| = \sum_{x \in P} v(x, c_i)$$

avec

$$v(x, c_i) = \begin{cases} 1 \text{ si } x \in c_i \\ 0 \text{ sinon} \end{cases}$$

L'histogramme ainsi obtenu, est une représentation en document d'un clip donné en utilisant l'approche "sac-de-mot". En faisant une analogie avec les documents textes, on dira que les centroïds correspondent à des "mots visuels". Le jeu d'apprentissage, composé uniquement de clip d'événements fréquents, devient à la fin de la présente étape, un ensemble d'histogramme d'occurrence (voir Fig 3.9). Cet ensemble qui servira d'entrée à l'algorithme LDA pour identifier de façon non-supervisée un ensemble de thèmes sémantiques qui traduira les différents sous-groupes d'événements normaux.

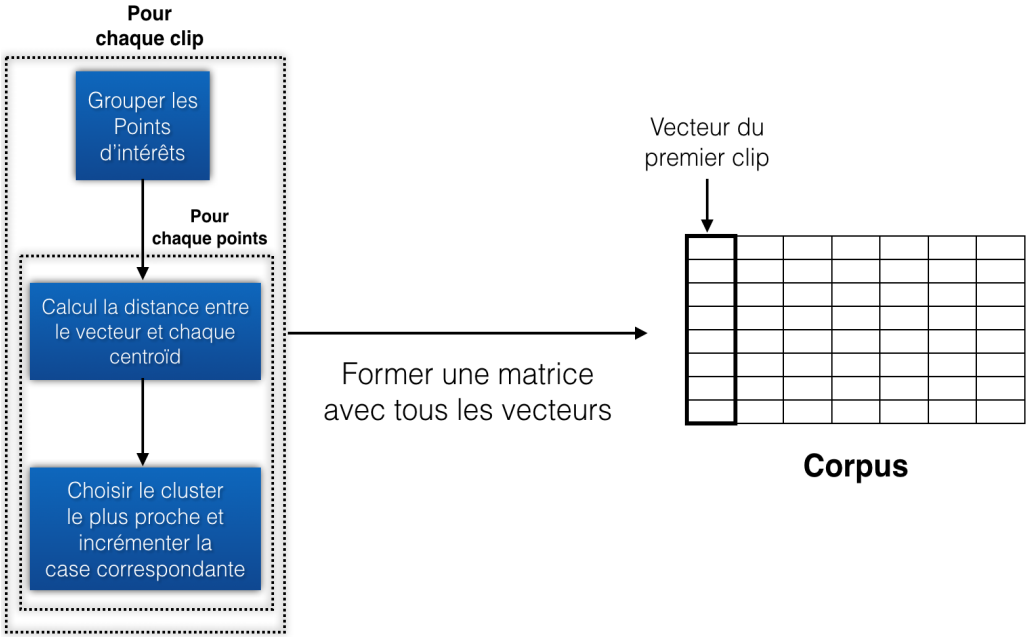


FIGURE 3.9 – Processus de génération du corpus

3.4.2 L'Allocation latente de Dirichlet : construction du modèle

Le modèle LDA a été introduit par Blei et al. [111] dans les années 2003 pour la modélisation des documents textuels. Le LDA est un modèle probabiliste génératif, utilisé pour décrire des documents de texte ou des données discrètes. On peut également le voir en tant que modèle bayésien hiérarchique à 3 couches (voir Fig 3.11) dans lequel chaque document est modélisé par un mélange de thèmes (topics), qui génére ensuite chaque mot du document. L'algorithme a connu un énorme succès surtout pour les tâches d'analyse de documents, les systèmes de recommandations d'article sur le web et bien d'autres applications. Avec son succès pour les documents textuels, il a été utilisé dans le domaine de la "vison par ordinateur" pour les tâches de classification de scènes naturelles [112] et même de modélisation d'événements [52].

La figure 3.10 présente de façon intuitive une compréhension du fonctionnement de l'algorithme. On peut remarquer à gauche, l'organisation structurelle des thèmes qui sont composés des mots du vocabulaire affectés d'une probabilité. Cela traduit la probabilité que le mot apparaîsse dans un document relatif au thème choisi. Pour un document donné, la distribution des thèmes est représentée par l'histogramme à droite dans l'image. Ainsi, les documents sont un mélange de thèmes. Autrement dit, un document peut être composé de plusieurs thèmes. Dans notre cas, cela se traduit par un mélange de groupe d'événements tous normaux. Les mots du document sont tirés à partir de cette distribution des thèmes.

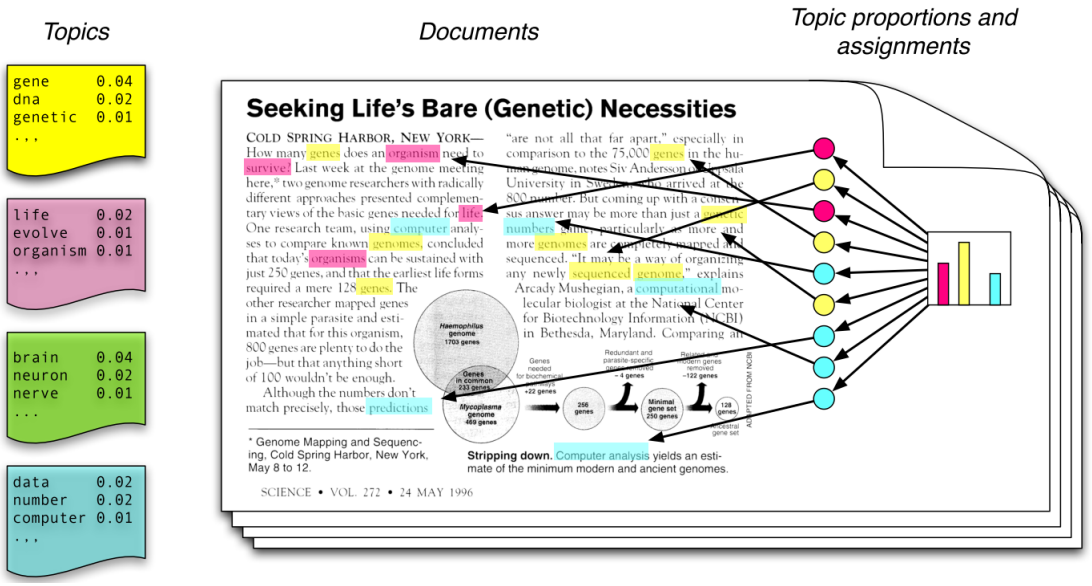


FIGURE 3.10 – Description schématique du LDA [113]

3.4.2.1 Paramètres et variables du modèle

La représentation graphique du modèle est donnée par la figure 3.11. A travers cette représentation, on peut observer les différents paramètres et termes du modèle. Les boîtes représentent des répliques du modèle qu'elles contiennent (par exemple, il y a N boîtes pour chaque document). Ces différents termes et paramètres peuvent être définis comme suit :

- **Le mot** : c'est la donnée discrète, l'élément primitif dans un dictionnaire. Il est représenté dans notre modèle par les centroïds issus de l'algorithme de clustering.

- **Le document** est formé d'un ensemble de mots. Pour les approches "sac-de-mots", l'ordre d'apparition des mots n'est pas important, mais le nombre d'occurence du mot dans le document doit etre considéré. On peut noter un document sous la forme : w = $(w_1, ..., w_N)$. Un document représente ici un clip.

- **Le corpus** est une collection de D documents. Dans une application de classification de documents, le corpus est composé de toute sorte de documents. Dans notre modèle, le corpus est représenté par l'ensemble des clips de la base d'apprentissage qui contient uniquement des vidéos d'événements normaux.

- $Z_{d,n}$ sont des variables qui désignent le thème choisi pour le mot $w_{d,n}$. Elles suivent une distribution multinominale.

- θ_d sont des paramètres qui expriment la distribution des thèmes pour un document. Il s'agit de l'histogramme de la figure 3.10. Ils suivent la loi de probabilité de Dirichlet.

- α et η sont respectivement les distributions a priori sur les paramètres θ_d et β. β_k représente la distribution du thème k.

Le modèle graphique nous renseigne également sur le niveau d'appartenance des paramètres dans le modèle ainsi que le nombre de fois que chaque paramètre doit être estimé. Les paramètres α et η sont estimés une seule fois pour l'ensemble du corpus. θ_d est estimé pour chaque document, β_k pour chaque thème et $Z_{d,n}$ pour chaque mot d'un document.

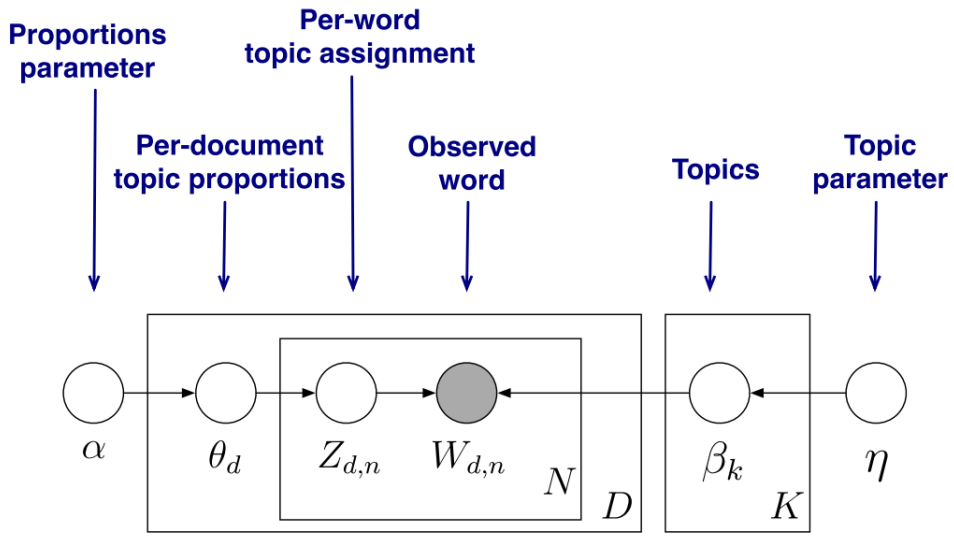


FIGURE 3.11 – Modèle graphique du LDA [113]

Etant donné le jeu d'apprentissage représenté dans la forme finale sous forme de documents, le processus génératif d'un document pour l'estimation des paramètres est le suivant :

Algorithme 4 : Processus génératif de LDA

Données : Corpus

1 Choisir $\theta \sim$ Dirichlet(α);

2 **pour** *les mots* W_n *dans le document* **faire**

3 Choisir un thème $Z_n \sim$ Multinominal(θ);

4 Choisir un mot $W_n \sim$ Multinominal(β_k), avec $k = z_n$;

3.4.2.2 Estimation des paramètres du modèle

Le modèle LDA est basé sur la loi de probabilité de Dirichlet en ce sens que pour générer un document, le tirage conditionnel des thèmes est fait suivant cette loi de probabilité. Les variables de θ de dimension k (k, le nombre de thème) tirées selon la loi de Dirichlet ont la propriété

suivante : $\theta_i \geq 0$ et $\sum_{i=1}^{k} \theta_i = 1$ et ont une densité de probabilité de la forme :

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \tag{3.8}$$

avec α un vecteur k-dimensions ($\alpha_i > 0$) qui contrôle l'homogénéité des θ_i et $\Gamma(x)$ est la fonction gamma.

La probabilité jointe du mélange de thèmes θ, des N thèmes z et de N mots w, étant donnés les paramètres α et β est donnée par l'équation :

$$p(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.9}$$

La loi marginale pour un document revient à la somme des probabilités jointes du mélange sur tous les thèmes θ :

$$p(\mathrm{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{3.10}$$

En observant le modèle graphique (Fig 3.11), on s'aperçoit que seuls les mots $W_{d,n}$ des documents sont les données observées, donc connues. Le reste des paramètres sont à estimer durant la phase d'apprentissage à travers l'inférence. Ainsi, les paramètres θ et z_n d'un document sont à déterminer à partir des mots qui le composent étant donnés les paramètres α et β. Différentes méthodes d'inférence sont utilisées dans la littérature pour résoudre le problème. On a par exemple les méthodes d'échantillonnage de Gibbs (collapsed Gibbs sampling) et les méthodes variationnelles telles que mean-field. Dans le cadre de nos travaux, la méthode variationnelle a été utilisée. Les paramètres α et β sont également inconnus mais peuvent être estimés grâce à l'algorithme EM (Expectation-Maximization). Le but de l'algorithme est de chercher le couple (α, β) qui maximise la log-vraissemblance (log-likelihood) du jeu d'apprentissage car l'équation 3.10 est impossible à évaluer. Les détails de l'algorithme sont donnés en annexe B.

$$l(\alpha, \beta) = \sum_{d=1}^{M} log\ p(\mathrm{w}_d|\alpha, \beta)$$

A l'issue de l'étape d'apprentissage, le couple (α, β) est connu. Pour un clip de test, en utilisant ce couple, il est possible de calculer la vraissemblance de ce dernier au regard du jeu d'apprentissage. En fixant un seuil sur cette vraissemblance, nous pouvons décider si le clip contient des événements rares ou normaux.

3.5 Résultats expérimentaux

Pour évaluer notre approche et la comparer aux approches existantes, nous avons utilisé la base de données publique qui traite de mouvement collectif de foule et abondamment utilisé dans la littérature. Dans cette section, nous allons présenter la base de données utilisée ainsi que les différentes expérimentations que nous avons effectuées.

3.5.1 Données d'évaluations

Etant donné que notre approche est plus orientée vers la détection globale d'événements, telle que les mouvements de foule, nous avons opté pour la base de données de l'université de Minnesota [114] relative à la détection de scène de panique de foule. Elle est constituée de onze vidéos de trois scènes différentes : une scène à l'intérieur d'un bâtiment et deux scènes à l'extérieur. Les vidéos présentent les mêmes chronologies dans le déroulement des événements, c'est-à-dire qu'elles commencent toutes par des personnes avec un mouvement aléatoire de marche et finissent par des mouvements de personnes qui courrent dans toutes les directions. L'objectif pour ce type de base est d'arriver à détecter de façon précise le moment où les personnes dans la scène se mettent à courir.

L'une des difficultés dans l'utilisation de cette base, pour une comparaison avec les méthodes de la littérature, est la constitution des jeux d'apprentissage et de test. Chaque auteur y va de sa composition, ce qui ne permet pas une comparaison objective. Dans le cadre des différentes évaluations que nous avons menées, les modèles sont construits en utilisant deux compositions différentes. La première composition dénommée "Comp 1" est constituée avec les séquences d'images d'événements normaux des vidéos de la scène intérieure pour l'apprentissage. Les évaluations sont faites sur le reste des vidéos de la base. Par cette composition, on peut facilement montrer que la méthode proposée est indépendante de l'environnement d'apprentissage. Pour la deuxième composition dénommée "Comp 2", nous avons pris les 300 premières images de chaque scène pour l'apprentissage et le reste pour le test.

FIGURE 3.12 – Exemple d'images de la scène sur le gazon : à gauche l'image contient un événement normal et à droite, elle contient un événement anormal

FIGURE 3.13 – Exemple d'images de la scène intérieure : à gauche l'image contient un événement normal et à droite, elle contient un événement anormal

FIGURE 3.14 – Exemple d'images de la scène sur la place publique : à gauche l'image contient un événement normal et à droite, elle contient un événement anormal

3.5.2 Evaluation quantitative

L'évaluation quantitative des méthodes de détection d'événements rares se fait de plusieurs manières. Plusieurs critères de performances existent dans la littérature pour montrer l'efficacité des différentes méthodes. Pour nos travaux, plusieurs tests ont été menés pour non seulement évaluer la performance de notre approche, mais également pour déterminer l'inflence des paramètres essentiels qui entrent en jeu dans sa mise en place. L'approche utilisée consiste à faire varier un paramètre parmi l'ensemble des paramètres tout en maintenant les autres fixes. L'un des critères de performance utilisé pour cette tâche est le AUC (Area Under Curve) de la courbe ROC qui met en évidence la capacité d'un système à bien prédire la classe d'une donnée de test en utilisant le "Taux de Faux Positif" (TFP) et le "Taux de Vrai Positif" (TVP). Pour rappel, la courbe ROC est tracée en utilisant les valeurs du couple (TFP, TVP) qui sont obtenues à partir de différentes valeurs de seuil de décision. Plus grande est la valeur de l'AUC, plus performante est le système de classification. La formule de calcul des valeurs de (TFP, TVP) est la suivante :

$$TFP = \frac{FP}{FP+VN} \text{ et } TVP = \frac{VP}{VP+FN}$$

avec :

FP : Faux Positif encore connu pour être les fausses alarmes ;

VN : Vrai Négatif. Les images correctement classées négatives ;

VP : Vrai Positif. Les images correctement classées positives ;

FN : Faux Négatif. Les images injustement classées comme négatives.

3.5.2.1 Influence de la combinaison des descripteurs de mouvement et d'apparence

Comme mentionné plus haut, nous avons combiné l'information d'apparence avec celle du mouvement pour l'amélioration de la détection en présence d'événements rares liés à l'apparence des entités qui les génèrent. Le but de l'expérience de cette section, est de voir la contribution réelle des descripteurs HOG et CHOP au côté du descripteur HOOF dans notre approche. La combinaison des deux types d'informations peut se faire suivant différents schémas de fusion : que ce soit directement au niveau des vecteurs de caractéristiques ou au niveau des scores de décision (fusion tardive). Pour cette expérience, nous avons opté pour la fusion au niveau des vecteurs de caractéristiques pour éviter d'avoir à construire deux modèles LDA avec deux différents corpus.

Pour fusionner les vecteurs de caractéristiques, nous procédons pour chaque point d'intérêt, à la concaténation du vecteur HOOF avec le vecteur du descripteur de forme. Nous avons fixé le nombre de mots à 20, le nombre de thème à 40, la valeur initiale de α à 0,3 et le taille des clips à 15. Les résultats obtenus sur la composition 1 de la base sont consignés dans le tableau 3.1.

Tableau 3.1 – Résultat de différentes combinaisons d'informations

Descripteurs	**HOOF**	**HOOF-HOG**	**HOOF-CHOP**
AUC	0,91	0,9165	0,9403

La meilleure combinaison de descripteur est celle avec CHOP. Les différentes valeurs des paramètres pour chacun des descripteurs de forme sont issues du chapitre 1. Vu la faible différence qu'il y a entre les performances, il ne serait pas possible de déclarer la supériorité de tel descripteur de forme sur un autre. Pour la suite des expérimentations, nous utiliserons la combinaison HOOF-CHOP.

3.5.2.2 Influence du nombre de thèmes pour le LDA

Le nombre de thèmes à découvrir par l'algorithme LDA représente le nombre de sous-groupe d'événements qu'on souhaite que l'algorithme découvre de manière non supervisée. Dans l'élaboration de notre approche, on peut définir un événement comme l'éxécution simultanée de plusieurs activités unitaires dans la même scène. Il sera intéressant d'analyser l'effet du nombre de thèmes sur les performances de la méthode. Nous avons fixé le nombre de mots visuels à 20, la taille des clips à 15 et la valeur initiale de alpha à 0,3. Les résultats obtenus sur la composition 1 de la base sont consignés dans le tableaux 3.2.

Tableau 3.2 – Résultat de l'influence du nombre de thèmes

Nombre de thèmes	10	20	30	40	50
AUC	0,9222	0,9260	0,9271	0,9403	0,9266

Il ressort du tableau que le nombre de thème influence sur les performances de l'approche. Cette influence reste néanmoins limitée. On peut conclure que ce paramètre influence très peu sur les performances, du moins pour cette base de test. En effet, avec la base que nous avons utilisé, le nombre d'activité unitaire est peu nombreux ce qui peut rendre peu sensible la méthode à la variation du nombre de thème.

3.5.2.3 Influence de la valeur initiale de α

Tout comme le nombre de thème, la valeur initiale du paramètre α peut influencer sur le modèle construit. Nous avons par le même processus évalué son influence réelle sur les performances de l'approche en variant sa valeur entre 0,1 et 0,9. Le nombre de mot est fixé à 20, le nombre de thème à 40 et la taille des clips à 15. Le tableau 3.3 récapitule les différentes performances

obtenues sur la composition 1 de la base . La meilleure performance est obtenue pour $\alpha = 0,3$.

Tableau 3.3 – Résultat de l'influence de la valeur initiale de α

α	0,1	0,3	0,5	0,7	0,9
AUC	0,9258	0,9403	0,9247	0,9247	0,9244

La différence entre les performances pour les différentes valeurs de α n'est pas grande, ce qui nous ramène au même cas que pour le nombre de thème. Ce paramètre influence également très peu sur les performances.

La figure 3.15 montre la courbe ROC associée au meilleur résultat parmi ceux présentés ci-dessus pour la composition 1.

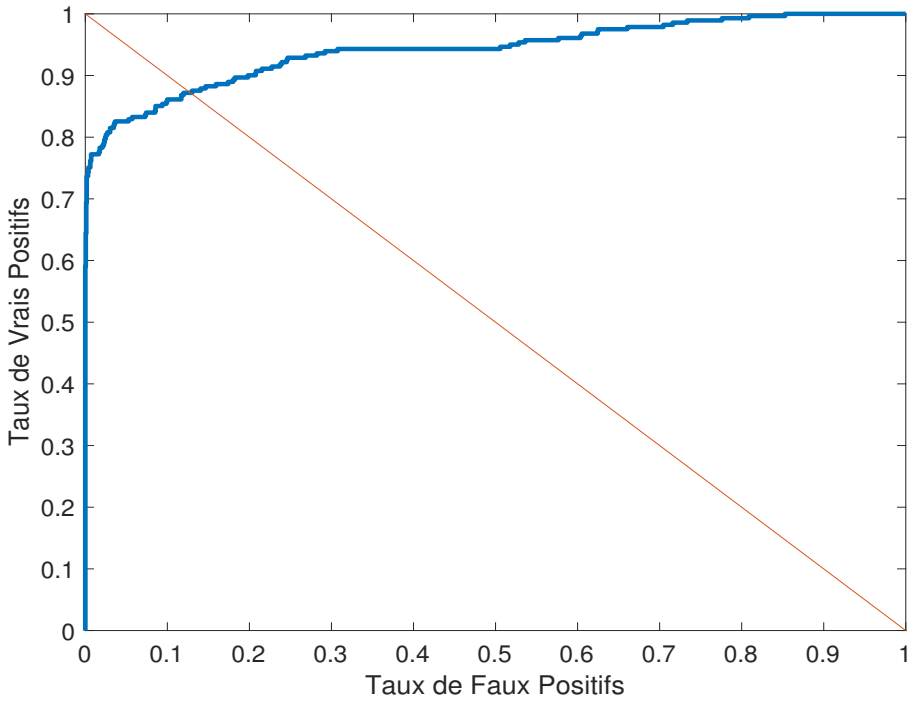


FIGURE 3.15 – Courbe ROC de la méthode BGMMAI sur la composition 1

3.5.2.4 Comparaison avec autres méthodes

Après l'évaluation de l'influence de différents paramètres clés de notre approche, il est important de nous comparer par rapport aux performances rapportées dans l'état de l'art. Comme nous l'avons déjà mentionné plus haut, une comparaison objective avec les approches de l'état de l'art s'avère difficile. Nous rapportons néanmoins dans cette section, quelques résultats même si la constitution des jeux d'apprentissage et de test sont différents. Notre approche peut être catégorisée parmi les approches "sac-de-mots". Quelques travaux ont utilisé la même approche sac-de-mots avec différents descripteurs et différents classifieurs. Nous rapportons dans le tableau 3.4 quelques résultats d'approches de cette catégorie.

Tableau 3.4 – Comparaison de la méthode avec les approches sac-de-mots [67]

Méthodes	SIFT	STIP	DT	Bag of Graph	**Notre Approche**
AUC	0,85	0,85	0,81	0,95	**0,9403**

On peut noter que notre approche dépasse ces méthodes à l'exception de la méthode basée sur les graphes. La différence entre les performances n'est pas énorme, ce qui nous amène à dire que notre approche est comparable à cette dernière. Dans la méthodologie, la différence entre notre approche et l'approche basée graphe se trouve essentiellement dans l'exploitation des points d'intérêt. Dans leur méthode, les points d'intérêt sont représentés sous forme de graphe avec pour sommet du graphe les dits points et pour arrête, la distance entre les vecteurs de descripteurs de chaque point. La construction du modèle d'événements est faite avec les graphes de la scène et l'algorithme SVM avec un noyau adapté aux graphes. Après la comparaison avec les approches de la même catégorie, nous présentons dans le tableau 3.5 les résultats de quelques autres méthodes. On retrouve les deux compositions de la base que nous avons retenues. Pour certaines méthodes qui utilisent la deuxième composition, le nombre d'images d'apprentissage peut varier, mais le principe reste le même.

Tableau 3.5 – Comparaison avec l'état de l'art

AUC / Méthode	**Gazon**	**Intérieure**	**Plaza**
Pure flot optique [52]	0,84		
Force sociale [52]	0,6		
NN [65]	0,93		
MDT temporel [81]	0,99		
MDT spatial [81]	0,97		
TCP [115]	0,988		
STCOG [66]	0,9362	0,7759	0,9661
HMOFP [116]	0,9976	0,9570	0,9869
Hajer et al. [117]	0,9872	0,9521	0,9934
MHOF [116]	0,9976	0,9570	0,9869
Notre Approche comp 1	0,9403		
Notre Approche comp 2	0,90	0,93	0,96

Les résultats montrent que notre approche bien que ne présentant pas les meilleurs résultats, a des bonnes performances et dépasse certaines méthodes de la littérature. Il aurait été intéressant de pouvoir comparer le temps de calcul des différentes méthodes. Mais malheureusement, la non disponibilité des codes des méthodes rend cette tâche difficile. La figure 3.16 montre les courbes ROC pour les différentes scènes avec la deuxième composition.

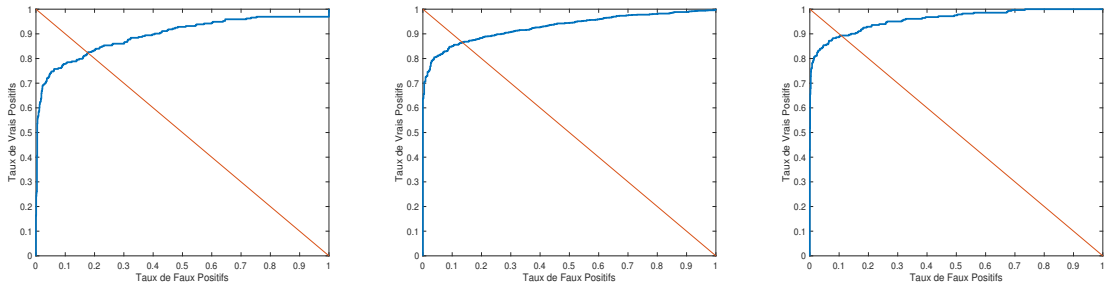


FIGURE 3.16 – Courbe ROC de la méthode BGMMAI sur la composition 2 :
De gauche à droite, les courbes ROC des scène 1, 2 et 3

3.5.3 Evaluation qualitative

L'évaluation qualitative consiste à visualiser les résultats obtenus en utilisant les meilleurs paramètres déterminés lors de l'évaluation quantitative de la sous-section 3.5.2 à partir du meilleur seuil déterminé sur la courbe ROC. Les figures 3.17 et 3.18 montrent les résultats de classification des images de la scène sur le gazon et plaza. Sur les images, le rectangle vert marque qu'il s'agit d'une image contenant des événements normaux tandis que la présence de rectangle rouge marque l'existence d'événement(s) rare(s) dans l'image. Ces résultats sont obtenus à partir du modèle d'apprentissage de la première composition de notre base et avec les paramètres suivants :

- Descripteurs : HOOF-CHOP ;
- Taille des clips : 15 images ;
- Nombre de mots et de thèmes : 40 ;
- Taille des fenêtres : 16 x 16 ;
- Valeur initiale de α : 0,3 ;

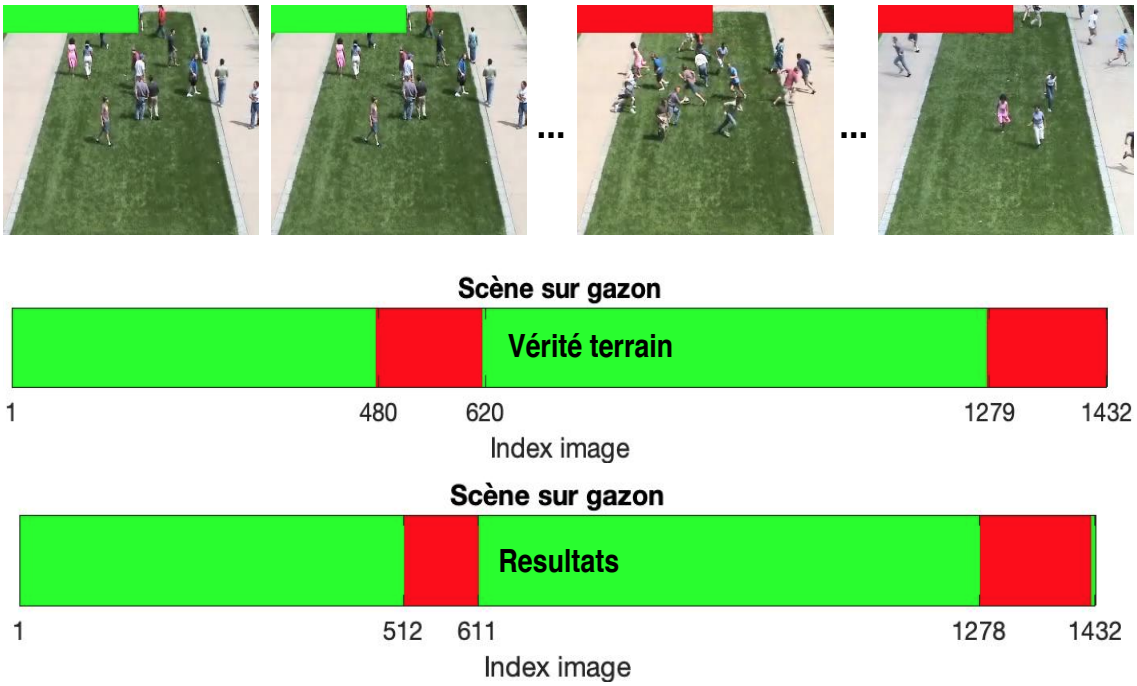


FIGURE 3.17 – Exemple de détection sur la scène de gazon

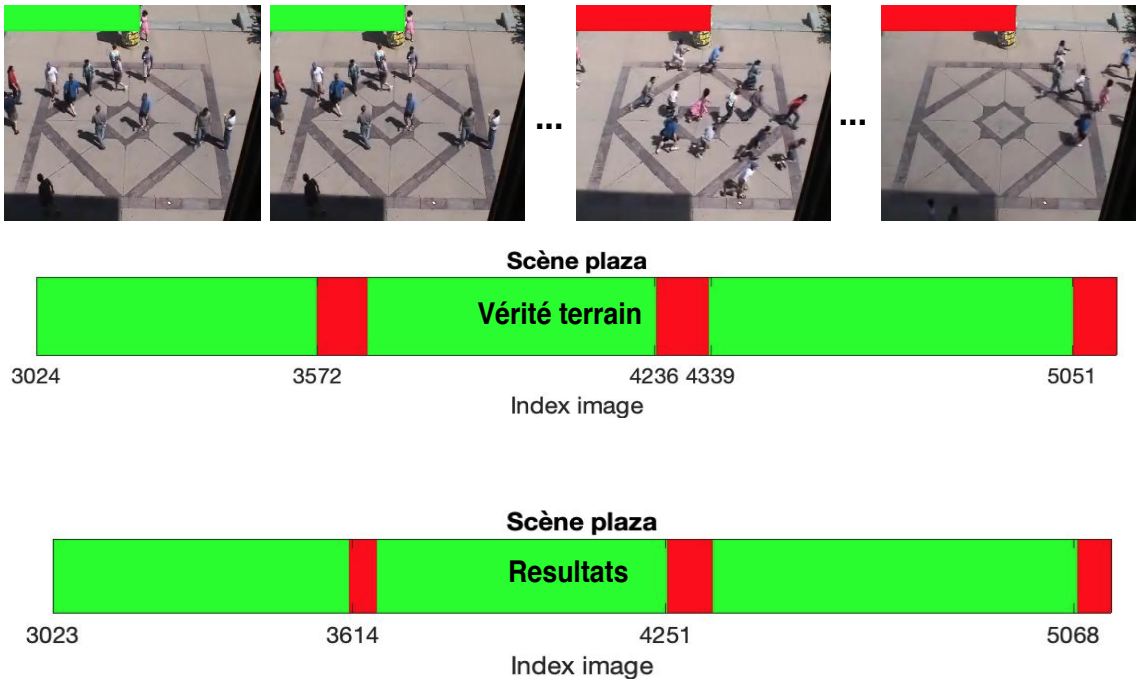


FIGURE 3.18 – Exemple de détection sur la scène extérieure place publique

On peut déduire de la comparaison des résultats avec les vérités terrains, que pour le seuil de décision optimal, la classification est satifaisante sur les deux scènes non apprises. En effet, sur l'ensemble des 3527 images des deux scènes, très peu sont mal classées. Le taux de bonne classification est de $Acc = 91,72\%$.

3.6 Conclusion

Dans ce chapitre, nous avons présenté la méthode BGMMAI, qui est une contribution pour la détection des événements globaux tel que les mouvements de foule dans une scène. La méthode se base sur la détection et le filtrage de points d'intérêt en utilisant la saillance visuelle. Les points d'intérêt ainsi filtrés permettent l'extraction d'informations pertinentes sur les entités saillantes présentes dans la scène. Ces entités sont par hypothèse les sources génératrices des événements. Deux types d'informations sont utilisés dans notre approche pour modéliser les événements. Il s'agit des informations de mouvement et d'apparence. L'information de mouvement est extraite à partir du flot optique et représentée par le descripteur HOOF tandis que l'information de forme est extraite grâce à différents descripteurs. La construction des modèles d'événements est faite grâce à l'algorithme LDA de découverte non supervisée de thèmes dans des corpus de documents. L'utilisation du LDA repose sur le fait qu'une scène est le siège du déroulement simultanée de plusieurs sous-événements que nous appelons "activités unitaires". Ainsi, à travers le LDA, nous avons pu modéliser les différents regroupement d'activités unitaires qui forment les événements normaux. Pour utiliser le LDA, nous nous sommes inspiré de l'approche "sac-de-mots" pour transformer les vecteurs de caractéristiques des points d'intérêt en des documents visuels nécessaires au fonctionnement du LDA. La classification des séquences d'images (clip) se fait à partir des scores de vraissemblance du document visuel qui représente le clip par rapport au modèle appris. La vraissemblance traduit en effet le dégré de ressemblance du document d'entrée à l'ensemble des documents du corpus d'apprentissage.

Nous avons également présenté dans ce chapitre les résultats expérimentaux de notre approche sur la base publique de l'université de Minesota. L'étude de l'influence de certains paramètres clés de l'algorithme LDA pour la création des modèles montre que ces derniers influencent très peu les performances de la méthode. En effet, avec la base de données utilisée, le nombre d'activité est peu nombreux. La recherche de la meilleure combinaison des descripteurs nous a permis de montrer que la combinaison HOOF-CHOP donnait les meilleures performances. La comparaison qualitative de nos performances avec celles des approches de la littérature montre des résultats équivalents prouvant la pertinence de l'approche.

Chapitre 4

Contribution à la détection et à la localisation d'évènements rares par analyse locale de scènes

> Le commencement de toutes les sciences, c'est l'étonnement de ce que les choses sont ce qu'elles sont
>
> ARISTOTE

Sommaire

4.1 Introduction

Dans le chapitre 3, nous avons présenté en détail notre méthode pour la détection d'événements rares globaux dans une scène. Pour cette catégorie d'approche, la détection se résume à l'identification du début et de la fin des mouvements collectifs brusques de la foule. Nous avons également présenté dans les chapitres précédents le fait que l'analyse de scène pour la détection d'événements rares concerne également les scènes contenant des événements locaux qui peuvent être difficile à détecter par une analyse globale de la scène. Différentes approches de l'état de l'art ont abordé la thématique sous cet angle. Une analyse globale de la scène mélange et noie les informations locales de ces événements avec celles des autres évènements. La détection est donc possible uniquement si le nombre d'événements rares présents dans la scène dépasse largement ceux normaux pour qu'une analyse globale de la scène permettent de les mettre en évidence. Un autre challenge vient de la variabilité intra-classe des événements rares et normaux. D'un côté, les événements normaux ne sont pas tous de même nature mais restent tout de même différents de ceux anormaux. Ces derniers présentent aussi des différences de nature en leur sein. De ce fait, il devient très difficile de détecter dans une même scène les différents types d'événements locaux.

Ce chapitre est consacré à la détection et à la localisation des événements rares locaux. Nous présenterons nos différentes contributions pour la détection de ces types d'événements rares. Ces contributions reposent essentiellement sur l'exploitation des informations de mouvements et d'apparence basées sur la saillance visuelle. Dans un premier temps, nous nous sommes penchés sur la détection de mouvements saillants dans une scène en exploitant les irrégularités des flots optiques de séquences d'images consécutives. Cette contribution permet dans le contexte de détection d'événements rares ou anormaux, de focaliser les méthodes d'extraction et de caractérisation d'événements sur des zones pertinentes de la scène. Dans un second temps, nous avons proposé une approche de modélisation locale utilisant comme descripteur les scores de saillance issus de l'étape précédente. Cette approche offre des résultats prometteurs sur des bases d'événements publiques abondamment utilisées dans la littérature.

4.2 Contribution à la détection de mouvements saillants dans une scène

La recherche exhaustive d'un certain nombre d'éléments dans une image reste l'approche la plus intuitive et basique. Il en est de même pour les applications de détection d'événements locaux. En effet, pour la création de modèle d'événements, on peut envisager l'extraction d'informations en tout point de la scène. Dans notre méthode BGMMAI, nous avons proposé une approche d'extraction de caractéristiques autour de points d'intérêt filtrés grâce à la saillance visuelle. Toutefois, la saillance visuelle utilisée dans cette approche focalise l'attention uniquement autour des objets saillants. La stratégie de filtrage utilisée repose sur la variance et la moyenne de la saillance de chaque point d'intérêt. A l'issue de l'étape de filtrage, les points d'intérêt retenus se retrouvent autour d'objets suffisamment saillants et non statiques.

L'idée de détection des mouvements saillants est de simplifier cette étape en supprimant l'étape d'extraction des points d'intérêt. Autrement dit, nous proposons de détecter les régions de la scène où se déroulent des mouvements irréguliers par rapport au reste de la scène. Notre hypothèse est que les événements rares générés par des entités en mouvement peuvent être assimilés à des mouvements irréguliers. Ainsi, nous avons proposé une méthode de détection de mouvements saillants, inspirée de celle de Hou et al. [118] pour la détection d'objets saillants dans une image. L'idée est de partir des informations d'orientation et de vitesse du flot optique, calculé entre deux images consécutives, pour mettre en évidence, à travers une reconstruction basée sur les signes de la transformée en cosinus discrète (TCD), les irrégularités possibles liées à ce dernier. Les étapes de notre approche sont résumées par la figure 4.1.

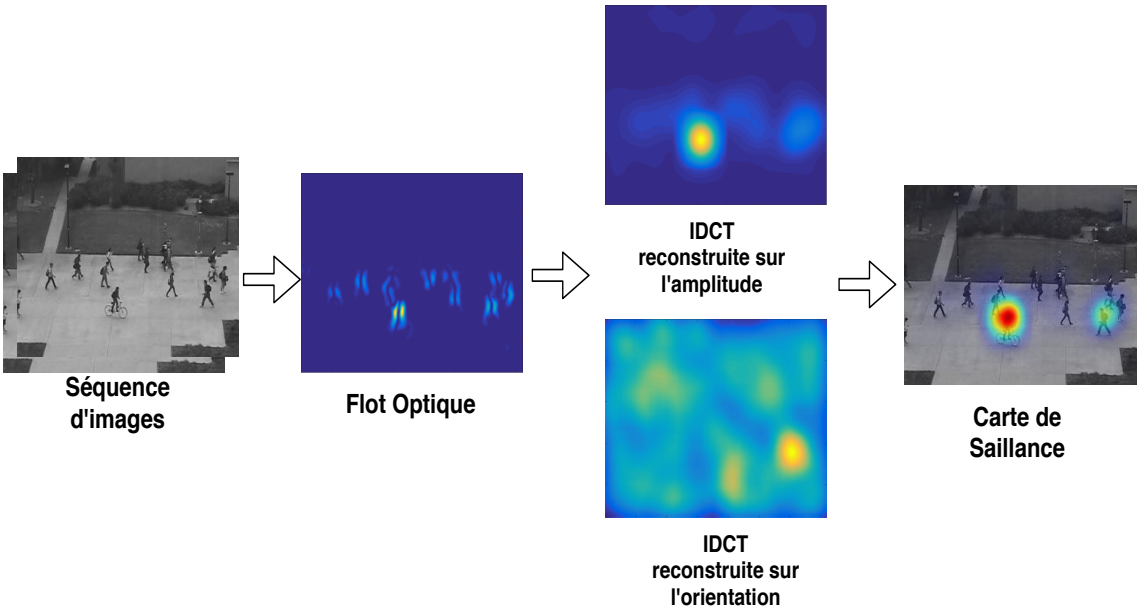


FIGURE 4.1 – Différentes étapes de la détection de mouvements saillants

La méhode que nous avons baptisée SF pour *Signature Flot* est simple, rapide à calculer, efficace et se résume en trois étapes. Elle commence par le calcul du flot optique entre deux images consécutives. Deux matrices sont obtenues à la sortie de cette étape et représentent les informations d'orientation et de vitesse au niveau de chaque pixel. Une transformée en cosinus discrète est appliquée séparément à chaque matrice. Les signes des coefficients de cette transformée sont très utiles pour la séparation des éléments réguliers et irréguliers. Comme pour la méthode de Hou et al, nous reconstruisons les matrices de départ en appliquant une transformée inverse sur les matrices contenant les signes des coefficients. Les résultats de ces reconstructions sont combinés pour générer une carte unique de mouvements saillants présents dans la scène.

4.2.1 Transformée en Cosinus Discrète : Rappel théorique

La Transformée en Cosinus Discrète (TCD) est utilisée dans beaucoup de domaines notamment en traitement d'images, en compression de vidéos, en tatouage d'images, etc. Elle est très

proche de la Transformée de Fourier à la différence de son noyau qui est un cosinus avec des coefficients réels. Elle est performante pour le regroupement d'énergie car l'information du signal est essentiellement portée par les coefficients basse fréquence. Cette propriété est un avantage majeur en compression d'images en ce sens que la reconstruction de l'image peut se faire à partir d'un petit nombre de coefficients non nuls sans grande perte d'information. De nombreux formats de compression de données se sont inspirés de la TCD. Pour une image, c'est la version 2-D de la TCD qui est généralement utilisée avec des blocs de taille 8×8. La TCD est une fonction linéaire inversible définie de $\mathbb{R}^N$ vers $\mathbb{R}^N$ et qui existe en trois versions en plus de l'inverse de la transformée. La TCD d'un signal 1-D noté $f(x)$ dans le domaine spatial est définie par l'équation 4.1 ci-dessous :

$$F(k) = \frac{1}{2}C(k)\sum_{x=0}^{7} f(x)cos\left(\frac{(2x+1)k\pi}{16}\right) \tag{4.1}$$

avec $F(k)$ la TCD dans le domaine fréquentiel et

$$C(m) = \begin{cases} \frac{1}{\sqrt{2}} & \text{si } m = 0 \\ 1 & \text{sinon} \end{cases}$$

Pour une image I représentée par la fonction d'intensité $f(x, y)$, la TCD 2-D peut être formulée selon l'équation 4.2 ci-dessous :

$$F(k,l) = \frac{1}{4}C(k)C(l)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y).cos\left(\frac{(2x+1)k\pi}{16}\right) cos\left(\frac{(2y+1)k\pi}{16}\right) \tag{4.2}$$

L'évaluation de l'expression brute de la TCD requière 8^4 opérations de multiplication. Cela induit une complexité importante et donc un temps de calcul conséquent. L'une des solutions couramment utilisée dans la littérature est la méthode de décomposition "row-column" qui ramène le calcul à deux opérations successives de décomposition, d'abord suivant les lignes puis suivant les colonnes. Cela permet de réduire considérablement le temps de calcul de la TCD 2-D. Des travaux dans la littérature se sont concentrés sur l'élaboration d'algorithmes pour la réduction du nombre d'opérations de multiplication. Ainsi, l'algorithme le plus efficace et le plus utilisé est celui de Loeffler et al. [119] qui, en se basant sur les techniques de "flow graph", ramène le calcul de la TCD à 11 opérations de multiplication pour la TCD 1-D. Récemment, Heyne et al. [120] ont proposé un algorithme de calcul de la TCD 1-D basé à la fois sur l'algorithme de Loeffler et l'algorithme CORDIC (COrdinate Rotation DIgital Computer) et qui n'utilise que 38 opérations d'addition et 16 opérations de décalage.

4.2.2 Construction des cartes de saillance

Par analogie à la signature d'une image proposée par Hou et al. [118], nous définissons comme signature de mouvement, les signes des transformées en cosinus discrète des cartes d'amplitude et d'orientation du flot optique. En effet, étant donné une séquence de deux images consécutives, l'extraction de l'information de mouvement donne pour chaque pixel deux composantes V_x et V_y. Ces deux composantes fournissent les informations d'orientation $\varphi_{x,y,t}$ et de vitesse $r_{x,y,t}$ en

chaque point (x, y) de l'image à l'instant t. L'estimation de ces deux composantes est faite par différentes méthodes que nous avons présentées dans le chapitre précédent.

$$\varphi_{x,y,t} = arctan\left(\frac{V_x}{V_y}\right) \tag{4.3}$$

$$r_{x,y,t} = \sqrt{V_x^2 + V_y^2} \tag{4.4}$$

La méthode que nous proposons estime les mouvements saillants à partir ces deux types d'informations. Autrement dit, nous recherchons les irrégularités aussi bien au niveau de la vitesse qu'au niveau des orientations. Soit $y = r + i \quad y, r, i \in \mathbb{R}^N$ une matrice représentant les orientations (ou la magnitude) du flot optique. y est le mélange d'informations régulières r et irrégulières i. r est supposé éparse et supporté par les bases spatiales standard tandis que i est assumé éparse et supporté par les bases de la TCD. Hou et al. dans leur article [71] ont démontré qu'il est possible d'isoler les informations redondantes d'une image en la reconstruisant à partir de sa signature basée sur le TCD. Partant de leur résultat, nous postulons qu'il sera possible d'isoler les mouvements réguliers donc dominants d'une scène, que ce soit en terme de vitesse ou d'orientation, en reconstruisant les matrices y à partir de leurs signatures ainsi définies.

$$Signature(y) = sign(DCT(y)) \tag{4.5}$$

Sur chaque signature, nous appliquons la transformée en cosinus discrète de type III encore appelée Transformée Inverse en Cosinus Discrète pour obtenir une carte reconstruite $\bar{y}$.

$$\bar{y} = IDCT(Signature(y)) \tag{4.6}$$

La carte des mouvements saillants est générée par filtrage du carré de la carte reconstruite suivant l'équation 4.7 :

$$S = g * (\bar{y} \circ \bar{y}) \tag{4.7}$$

Où g est un filtre gaussien. Un tel filtrage est nécessaire pour éliminer d'éventuels bruits et améliorer la qualité de la carte de saillance.

A cette étape, nous obtenons des cartes des mouvements saillants en terme de vitesse S_r d'une part et en terme d'orientation S_φ d'autre part. Pour prendre en compte les informations des deux cartes et obtenir une carte de saillance unique, nous procédons à la multiplication de S_φ par S_r. Cela permettra de se focaliser uniquement sur les mouvements saillants induits aussi bien par la vitesse que par l'orientation.

$$S_f = S_\varphi \circ S_r \tag{4.8}$$

La carte S_f est normalisée pour obtenir des valeurs entre 0 et 1 avec 1 pour très saillant et 0 pour le contraire.

4.2.3 Evaluation de la méthode

4.2.3.1 Evaluation quantitative

Nous avons évalué la méthode sur la base de données publiques UCSD Ped2 d'événements anormaux. Cette base est dédiée à la détection des événements rares ou anormaux. Elle contient des événements tels que le passage d'un cycliste, d'une voiture, d'une personne sur trotinette avec des mouvements "zig zag", horizontaux ou obliques. La scène est dominée par le passage de piétons avec des vitesses similaires. La base est divisée en une série de 16 vidéos d'apprentissage et de 13 vidéos de tests. Nous nous intéressons uniquement aux vidéos de test, plus précisément aux séquences où les mouvements irréguliers apparaissent. Pour ce faire, nous avons produit des vérités terrain complémentaires à celles fournies pour la détection des événements rares. En effet, un mouvement saillant n'est pas forcément un mouvement anormal. Pour cette base précisément, nous retrouvons des mouvements saillants tels que des personnes qui marchent à contre-sens par rapport au reste des piétons mais qui ne sont pas pris en compte dans les vérités terrain initialement produites. Ces types de mouvements saillants ont été pris en compte pour mener une évaluation judicieuse. Il a été alors question de refaire une labélisation binaire de carte de vérité terrain en rajoutant dans les cartes initiales les mouvements irréguliers non anormaux. Plusieurs critères de performances peuvent être utilisés dans le cadre des évaluations quantitatives. Nous avons mené une étude comparative entre notre méthode et celle de Loy et al. [71] en utilisant comme critères le F_2Score, l'accuracy et le coefficient de Dice. Le F_2Score est une mesure d'efficacité d'un système de classification binaire qui prend en compte la précision et le rappel tout en donnant plus de poids à la précision. L'accuracy permet d'avoir une idée plus globale du taux de bonne détection que ce soit pour la classe positive que pour la classe négative. Le coefficient de Dice est utilisé pour mesurer la similarité entre deux masques, l'un représentant la vérité terrain et l'autre le résultat de la segmentation. D'autres métriques aussi intéressantes, telque NSS et le coefficient de Pearson, existent et peuvent être utilisés pour cette tâche.
La méthode de Loy et al. [71] est basée sur la méthode du résidu spectral calculé par la transformée discrète de Fourier. Dans la littérature, les méthodes de saillance visuelle se limitent en général à la détection d'objets saillants dans des images statiques. Loy et al. ont, dans leur étude, pointé du doigt un manque d'attention pour la thématique de la détection de mouvements saillants. Quoique proche de la détection d'objets saillants, la détection de mouvements saillants est plus qu'un ajout d'informations temporaires dans la détection d'objets saillants. En effet, des travaux ont abordé la détection d'objets saillants temporaires qui consiste à faire une détection d'objets saillants dans une fenêtre temporaire. Nous avons choisi de nous comparer à la méthode de Loy et al. car c'est celle qui se rapproche le plus du contexte de notre étude.

Avant de présenter les résultats quantitatifs, il est important de donner une définition des différentes mesures effectuées et qui ont servi au calcul des différents scores. Un pixel est déclaré appartenant à un mouvement saillant si son score de saillance est supérieur à un seuil prédéfini. Etant donné une carte binaire comme vérité terrain et une autre obtenue après seuillage de la carte de saillance, nous définissons comme :

- Vrai Positif (VP) : les pixels qui sont simultanément saillants au niveau de la vérité terrain et de la carte de saillance ;

- Vrai Négatif (VN) : les pixels non saillants qui ont été détectés comme tel ;

- Faux Positifis (FP) : les pixels non saillants qui été détectés comme saillants ;

- Faux Négatifs (FN) : les pixels saillants non détectés.

Ces indicateurs sont calculés (voir équations 4.9, 4.10 et 4.11) au niveau de chaque séquence d'images. La performance globale sur une vidéo et sur la base d'évaluation est une moyenne des performances individuelles au niveau de chaque image.

$$Rappel = \frac{VP}{VP + FN}, \quad Precision = \frac{VP}{VP + FP} \tag{4.9}$$

$$Accuracy = \frac{VP + VN}{VP + FN + FP + VN}, \quad F_2Score = 5 * \frac{Precision * Rappel}{4 * Precision + Rappel} \tag{4.10}$$

$$Dice = \frac{2 * VP}{2 * VP + FN + FP} \tag{4.11}$$

Pour effectuer une comparaison judicieuse des performances des deux méthodes, il est important de déterminer les valeurs optimales de l'écart-type σ du filtre gaussien et du seuil de décision. En effet, le filtre gaussien appliqué à la carte de saillance influence la performance des méthodes. Pour éviter que les mouvements saillants détectés ne coïncident pas bien avec la réalité, il est important de bien choisir la valeur de ce paramètre. Le seuil influence également les performances. Pour ce faire, nous avons étudié l'influence de ces deux paramètres en utilisant le coefficient de Dice comme indicateur de performance. La valeur du seuil évolue dans l'intervalle $[0, 1]$ et celle de sigma dans l'intervalle $[0, 0.1]$.

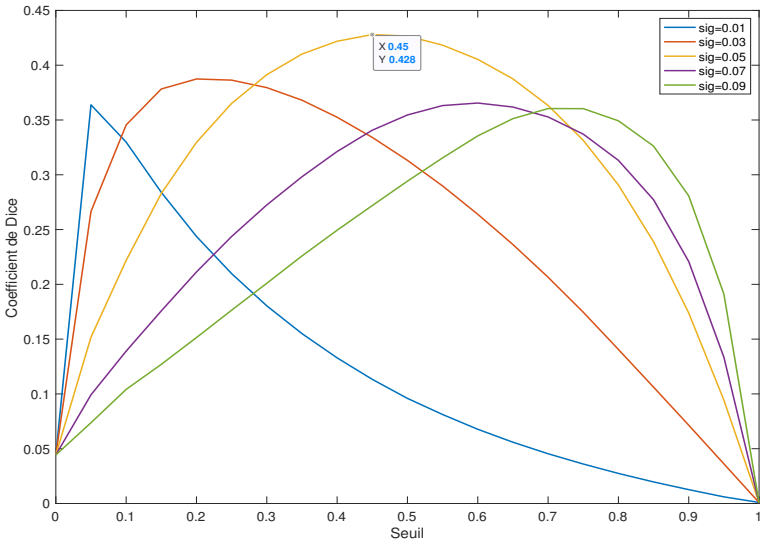


FIGURE 4.2 – Performance de notre méthode en fonction du seuil et de σ

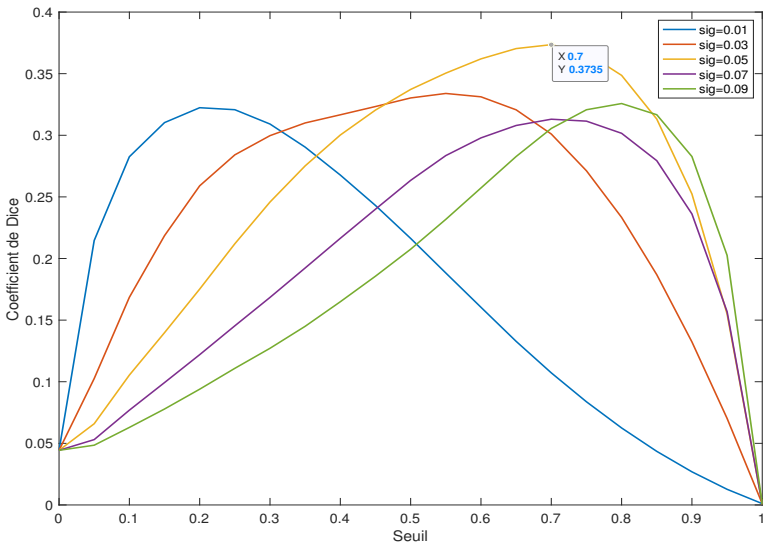


FIGURE 4.3 – Performance de la méthode de Loy et al. fonction du seuil et de σ

Les figures 4.2 et 4.3 montrent respectivement l'évolution du score de Dice en fonction du seuil de décision pour notre méthode et celle de Loy et al. [71]. Les différentes courbes nous permettent de mesurer l'inflence des deux paramètres sur les performances des deux méthodes. Elles nous permettent également de déterminer les valeurs optimales de ces derniers. Ainsi, le couple optimal $(seuil, \sigma)$ pour notre méthode est de $(0.45, 0.05)$ et de $(0.7, 0.05)$ pour celle de Loy et al. avec des scores respectifs de $0,43$ et $0,37$.
Afin d'analyser de plus prêt les différentes performances, nous avons fixé les deux paramètres aux valeurs optimales pour chaque méthode afin de calculer les précisions et rappels ainsi que l'accuracy de ces dernières. Les résultats sont consignés dans le tableau 4.1.

Tableau 4.1 – Résultats comparatifs des deux méthodes

	Méthode de Loy et al. [71]	Notre Méthode
Acc	97,41%	96,69%
Rappel	41,93%	40,46 %
Precision	42,98%	57,98%
F_2Score	0,4214	0,4305

Au vu de ces résultats, nous pouvons confirmer la supériorité de notre approche en considérant comme indicateur le F_2Score. Néanmoins, les valeurs de rappel et de précision montrent que notre méthode est moins bonne en terme de rappel quoi que sensiblement égale mais présente une meilleure précision. L'importance qu'on accorde à la précision ou au rappel dépend de l'application visée par la méthode. Dans notre cas, une méthode de détection de mouvements saillants peut être utilisée en amont pour réduire le champ de recherche des événements rares à

travers une proposition de régions candidates. Il nous semble donc important de proposer toutes les régions contenant des événements rares tout en réduisant au maximum les faux positifs. En partant de ces contraintes, la meilleure méthode serait celle proposant peu de régions candidates avec le maximum de vrais positifs et le minimum de faux positifs. Les résultats montrent sur cette base que notre méthode propose peu de régions candidates compararée à celle de Loy et al. mais avec une meilleure précision.

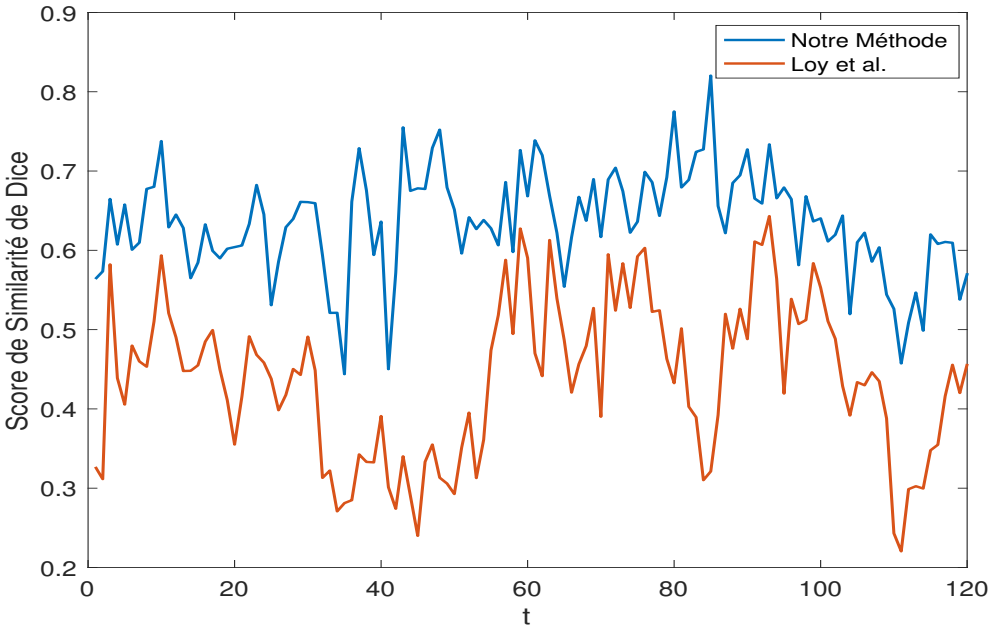


FIGURE 4.4 – Evolution du coefficient de Dice pour la vidéo 5

Il faut garder à l'esprit que les résultats des performances observées sont des moyennes de performances au niveau de chaque séquence d'images des vidéos testées. L'observation (voir figure 4.4) de l'évolution du coefficient de Dice pour la vidéo 5 par exemple, qui correspond au passage d'un cycliste dans la scène, montre une variation en dent de scie de la performance pour chaque séquence. Il est facilement observable qu'en fonction des mouvements dans la scène, il arrive que l'algorithme se trompe pour se focaliser sur d'autres mouvements qui ne sont pas en réalité saillants. On peut noter par exemple pour un mouvement comme le passage d'un cycliste ayant une couleur proche de l'arrière-plan de la scène, que les méthodes peuvent ne pas continuer à détecter le cycliste. Cela se justifie par le fait que les informations du flot optique dans cette zone et à cet instant seront presque nulles. Le faible pourcentage de précision et de rappel est dû au fait que les entités qui génèrent des mouvements saillants dans la scène sont généralement de petites tailles et donc représentées par un nombre très petit de pixel comparé au nombre restant de pixel de la scène. L'évaluation étant réalisée au niveau du pixel, le faible pourcentage est principalement dû aux mouvements de piétons ou de cyclistes.

4.2.3.2 Evaluation qualitative

Les résultats qualitatifs permettent de visualiser concrètement la détection des mouvements saillants. Les résultats expérimentaux présentés ici sont obtenus en utilisant l'algorithme de Horn-Shrunk pour le calcul du flot optique et en fixant σ à sa valeur optimale déterminée dans la section précédente. La performance de la méthode est tributaire de la performance de l'algorithme de calcul du flot optique en ce sens qu'elle se base uniquement sur les informations de mouvement fournies en entrée. La figure 4.5 présente quelques résultas qualitatifs issus d'extrait vidéo du

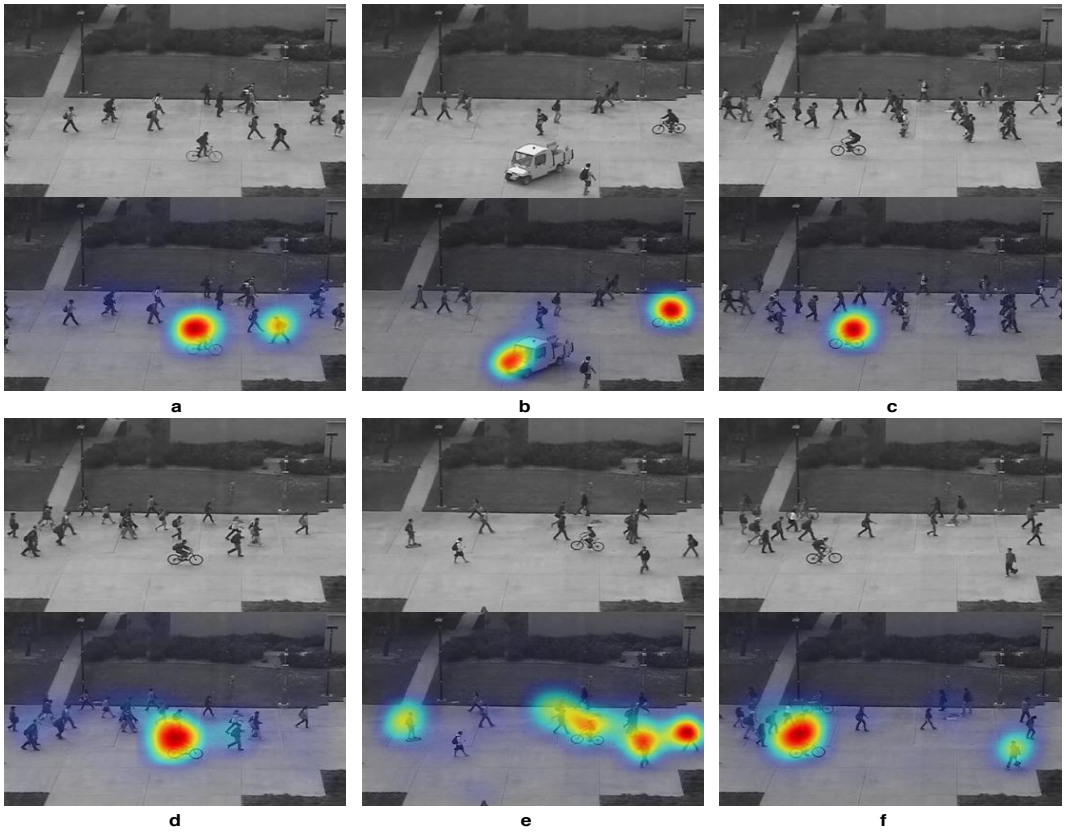


FIGURE 4.5 – Résultats qualitatifs de détection de mouvements saillants (Cyclistes (a-c-d-e-f), voiture (b) ou direction incohérente (a-e-f)

jeu de test. Pour chaque paire d'images (a, b, c, d, e et f), l'image en haut correspond à l'image originale à l'instant t et l'image en bas correspond à la carte de saillance représentée par une carte de chaleur superposée sur l'image originale. Ces quelques résultats montrent une bonne détection des mouvements saillants par notre méthode. Ainsi, le résultat a montre que la méthode focalise réellement l'attention sur le cycliste et la personne qui marche à contre-sens. Le résultat b montre une focalisation sur la voiture et le cycliste. Néanmoins, il arrive que la méthode oublie un mouvement saillant ou détecte des mouvements qui ne le sont pas. On peut voir sur le résultat b, deux personnes avec un mouvement à contre-sens non détectées. Comme on peut le noter sur l'image b, il arrive parfois que la zone déterminée comme étant une zone de mouvement saillant ne couvre pas tout l'objet générant le mouvement. Ainsi, on peut voir que seule une partie de la voiture est mise en évidence. Cela explique également le faible pourcentage de précision que nous pouvons observer au niveau de l'évaluation quantitative. En supposant que la mise en évidence

d'une partie de l'objet générant le mouvement saillant peut être considérée comme étant une bonne détection, nous avons ajusté les vérités terrain pour chaque détection en éliminant les parties non détectées de l'objet. Cela a l'avantage d'avoir une mesure plus réaliste de la précision des algorithmes. En procédant ainsi, nous avons obtenu une précision de $76,12\%$ pour notre méthode contre $65,87\%$ pour celle de Loy et al.

4.3 Détection et localisation d'événements rares par analyse de mouvements saillants

4.3.1 Méthodologie

Au-delà de la détection de mouvements saillants pour la proposition de régions de recherche d'événements rares, serait-il possible d'apprendre un modèle d'événements à partir des cartes de saillance ? Quel serait l'apport des descripteurs présentés dans les chapitres précédents dans une approche de modélisation locale ? Telles sont les questions qu'aborde l'étude que nous présentons dans cette section. En effet, il sera question ici d'étudier, à partir de la méthode de la figure 4.6, la performance de l'utilisation des scores de saillances comme descripteur et de la comparer à l'histogramme des orientations du flot optique (HOOF). L'étude de la fusion des descripteurs de mouvements et des différents descripteurs de formes sera également faite.

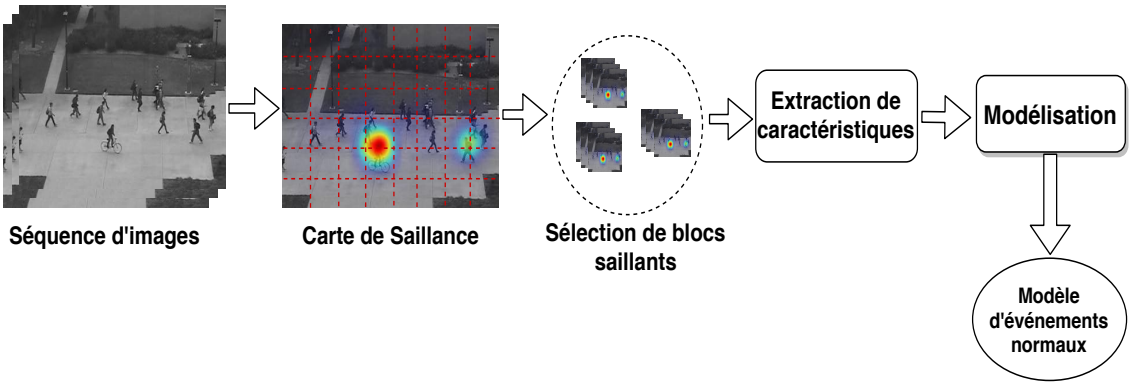


FIGURE 4.6 – Schéma du déroulement de la méthode

4.3.1.1 La sélection des blocs

Dans cette approche, nous commençons par estimer la carte des mouvements saillants de la scène pour une séquence d'images d'entrée. Il s'ensuit une subdivision de la scène en des blocs de taille égale sans chevauchement parmi lesquels les plus saillants seront retenus pour la phase d'extraction de caractéristiques. A défaut de fixer un seuil sur le score de saillance pour la sélection, nous avons choisi un seuil dynamique th pour chaque image. Ce seuil est empiriquement fixé à deux fois la moyenne des scores de saillance de l'image. Les blocs retenus sont ceux dont le score moyen est supérieur à th. Il est important de rappeler que contrairement à la section précédente, nous cherchons ici à classer les mouvements en fonction de leur nature et non juste à mettre en évidence les mouvements saillants. L'hypothèse sur laquelle se fonde notre démarche est qu'un mouvement saillant n'est pas forcément un événement rare mais le contraire est vérifié.

Durant la phase d'apprentissage, les vecteurs de caractéristiques sont extraits dans tous les blocs saillants de toutes les vidéos de la base d'apprentissage. Une étape supplémentaire d'élimination des informations redondantes peut être envisagée pour accélérer le temps de calcul. Le modèle d'événements normaux est appris grâce à la variante "one class" de l'algorithme populaire SVM. Pour une séquence d'images de test, le même procesus est appliqué jusqu'au niveau de l'extraction des descripteurs. Une fois les caractéristiques extraites, la phase de prédiction permet d'obtenir un score de classification pour chaque bloc de l'image. Une image est déclarée ne contenant aucun événement rare si le score maximum de ses blocs est inférieur au seuil.

4.3.1.2 Les descripteurs mis en jeu

Dans le même esprit que les travaux sur la méthode BGMMAI, deux catégories de caractéristiques seront mises en jeu pour voir dans quelle proportion elles contribuent à l'amélioration des performances de l'approche. Ainsi, dans la catégorie des descripteurs de mouvements, nous avons :

- l'histogramme des orientations du flot optique (HOOF). Traditionnellement utilisé dans bon nombre d'approches pour la description des informations de mouvements dans une scène. Nous l'avons présenté en détail au niveau du chapitre 3 avec la méthode BGMMAI,

- les scores de saillance des pixels des blocs retenus. En effet, pour chaque bloc retenu après filtrage, un vecteur de saillance sera associé et qui n'est rien d'autre qu'une vectorisation de la matrice de score du bloc.

$$descripteurSaillance = vect(score(bloc[i]))$$

Le recours à cette information pour la classification des mouvements est une suite logique des travaux présentés dans la section précédente. Dans la phase précédente, un score de saillance est attribué à chaque pixel au regard du mouvement global dans la scène. Une observation de l'évolution des scores non normalisés montre qu'il existe une corrélation entre le score de saillance et la nature du mouvement saillant. Une modélisation basée sur les scores des blocs saillants peut permettre de distinguer un mouvement saillant ordinaire d'un mouvement saillant rare.
Dans la catégorie des descripteurs de forme, nous avons retenu :

- l'histogramme des orientations de gradient (HOG),

- l'histogramme des orientations de phase (CHOP),

- les motifs binaires locaux (LBP).

Les détails sur ces descripteurs sont donnés dans le chapitre 1. L'idée de rajouter des descripteurs de forme est d'une part de prendre en considération les événements rares qui seraient liés à la forme des entités présentes dans la scène et d'autres part d'évaluer leur contribution effective à la détection des événements. Différents schémas de fusion des différentes informations peuvent être envisagés. Dans la section d'évaluation, nous présenterons les schémas de fusion mis en oeuvre pour l'évaluation de la méthode.

4.3.1.3 Description du SVM One class

Le choix de l'algorithme SVM trouve son fondement dans sa popularité dans l'état de l'art pour la modélisation pour des problèmes de classification. Plus encore, l'algorithme dispose d'une version monoclasse généralement appelée "one class" qui s'adapte très bien à la problématique d'apprentissage d'une seule classe de données et de détection de valeurs aberrantes (outliers). En effet, notre approche peut être assimilée à la détection de valeurs aberrantes car nous ne disposons que des données sur les événements fréquents. L'objectif est de construire un modèle qui définisse une frontière autour des vecteurs de caractéristiques des mouvements fréquents. Ces vecteurs constituent notre monoclasse pour l'entraînement du modèle. Dans la phase de prédiction, tous les vecteurs de caractéristiques qui sont au-delà de la frontière seront déclarés rares. L'idée principale d'un SVM one class est justement de trouver l'hyperplan de marge maximale à l'aide d'une fonction à noyau appropriée pour mapper la plupart des données d'entraînement vers un seul côté de l'hyperplan. Ce SVM maximise la distance de cet hyperplan par rapport à l'origine. Il peut être considéré comme un SVM standard à deux classes dans lequel toutes les données d'apprentissage se trouvent d'un seul côté de l'espace des fonctions, tandis que les points de données aberrantes se trouvent de l'autre côté. Pour trouver l'hyperplan de la marge maximale, le problème sous-jacent du SVM one class est formulé comme le problème quadratique suivant :

$$\min_{w,\rho} \frac{1}{2}||w||^2 + \frac{1}{vN}\sum_{i=1}^{N}\zeta_i - \rho \tag{4.12}$$

tel que :

$$w^T.\Phi(\times_i) \geq \rho - \zeta_i, \zeta_i \geq 0 \tag{4.13}$$

où $\times_i$ pour $i = 1, 2, ..., N$ est l'ensemble des données d'apprentissage et w est le vecteur de poids appris. De plus, ρ est le décalage et le paramètre prédéfini $v \in (0, 1]$ représente une limite supérieure sur la fraction de données que l'on peut localiser du côté des données aberrantes. $\Phi(\times_i)$ est la fonction de projection du vecteur de caractéristiques $\times_i$ dans un espace de grande dimension F. Comme pour le SVM classique, cette fonction de projection peut être définie implicitement en introduisant une fonction à noyau comme nous l'avons présenté dans le chapitre 1.

4.3.2 Evaluation de l'approche

4.3.2.1 Jeux de données

Pour évaluer la méthode, nous avons conduit des évaluations niveau frame et niveau pixel qui consistent à classer les séquences d'images, en fonction des scores de classification des vecteurs de caractéristiques des blocs saillants, en des séquences normales ou anormales. Nous avons utilisé les deux bases de test UCSD [1] Ped 1 et Ped 2. La base UCSD Ped 1 est constituée de 34 vidéos d'apprentissage et de 36 vidéos de test de résolution 158×238 pixels avec une durée fixe de 200 images. La deuxième base UCSD Ped 2 est celle utilisée dans la section précédente. Toutes les vidéos d'apprentissage contiennent uniquement un nombre varié des piétons qui marchent dans la scène. Les événements rares sont assimilés aux apparitions de cyclistes, de personnes sur

1. http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

trottinette et de voitures avec des mouvements à orientations diverses. La tableau 4.2 résume les anormalités telles qu'énumérées dans [81]. Les données du tableau sont sous la forme a/b qui équivaut à $nombre\_clips/nombre\_anomalies$. Il faut noter que les clips contiennent une ou plusieurs anomalies. Les vérités terrain avec des labélisations niveau frame sont fournies pour

Tableau 4.2 – Composition de la base UCSD

Base	Bicyclette	Patineur	Chariot	Marcher à travers	Autres	Total
Ped1	19/28	13/13	6/6	3/4	3/3	36/54
Ped2	11/19	3/3	1/1	0/0	0/0	12/23

toutes les vidéos des deux jeux de données. La labélisation niveau pixel qui permet une évaluation niveau pixel est fournie pour certaines vidéos de la base UCSD Ped 1 et pour toutes les vidéos de la base Ped 2.

4.3.2.2 Résultats Expérimentaux

Nous avons procédé dans cette sous-section à deux types d'évaluation des performances de l'approche. Il s'agit de l'évaluation niveau frame et l'évaluation niveau pixel. Pour l'évaluation niveau frame, une image est classée anormale, si un des blocs a un score supérieur au seuil de décision. Pour ce faire, après avoir calculé le score de tous les blocs de l'image, nous retenons comme score final de l'image, le score maximal de tous les blocs. Pour l'évaluation niveau pixel, le protocole suivi est celui utilisé par la plus part des travaux de la littérature. A partir des scores des blocs, un masque binaire est généré pour un seuil prédéfini. Ce masque est comparé aux masques de la vérité terrain disponible. Une image est déclarée contenant un événement rare, si le masque généré couvre au moins 40% de la vérité terrain. A partir de ces protocoles, pour chaque jeu de données, nous calculons l'aire sous la courbe ROC (AUC), le EER (Error Equal Rate) et le RD (Rate Detection) en variant le seuil de décision. Ces indicateurs de performances sont les plus utilisés dans la littérature. Le EER est calculé à partir des coordonnées (TFP, TVP) du point d'intersection de la courbe avec la diagonale décroissante de son plan de tracé. Autrement dit, le point où $FPR = 1 - TPR$. L'EER est utilisé pour l'évaluation niveau frame tandis que le $RD = 1 - EER$ est utilisé pour l'évaluation niveau pixel.

A - Etude de l'influence du schéma d'extraction

Une première expérimentation a été menée avec des caractéristiques extraites dans les blocs ou volume de blocs et un noyau linéaire pour l'algorithme SVM. En effet, aussi bien pour l'apprentissage que pour le test, les descripteurs sont extraits au niveau spatial dans chaque bloc. Pour l'extraction dans les volumes de blocs (information spatio-temporelle), une combinaison des informations spatiales est faite soit par concaténation soit par calcul de la moyenne des vecteurs de caractéristiques spatiaux. Avec un noyeau linéaire pour le SVM, le temps de calcul pour la construction du modèle et pour le test est fortement réduit. Le tableau 4.3 présente des résul-

Tableau 4.3 – Evaluation niveau frame avec le noyeau Linéaire

	Ped 1		Ped 2	
Méthodes	**AUC**	**EER**	**AUC**	**EER**
Saillance-Spatiale	0,78	0,28	0,79	0,30
Saillance-SpatioTemporelle	0,77	0,30	0,80	0,30
HOOF-Spatiale	0,77	0,30	0,75	0,33
HOOF-SpatioTemporelle	0,76	0,32	0,77	0,33

tats obtenus après l'évaluation niveau frame en utilisant le score de saillance et l'histogramme des orientations du flot optique (HOOF) comme descripteur. Au regard de ces résultats, on peut observer qu'en utilisant le score de saillance comme descripteur, nous obtenons des performances similaires à celles de HOOF. Le descripteur de saillance présente une légère avance sur le descripteur HOOF pour l'ensemble des deux bases. Ces résultats nous montrent également que l'utilisation de l'information spatio-temporelle telle qu'effectuée ici ne présente pas un réel avantage sur l'information spatiale. Les courbes ROC associées à ces résultats sont présentées à la Figure 4.7.

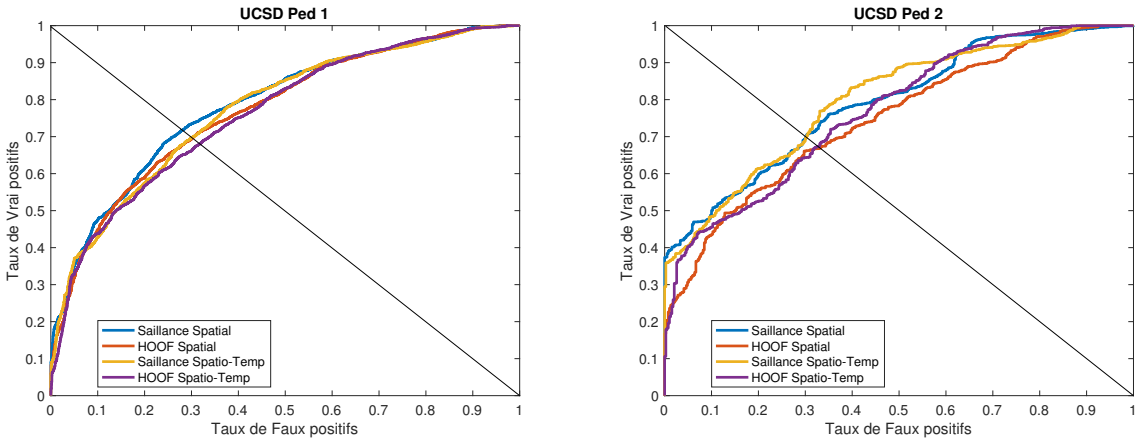


FIGURE 4.7 – Courbes ROC : Evaluation niveau frame

En complément de ces résultats, nous avons procédé à l'évaluation niveau pixel dont les résultats sont consignés dans le tableau 4.4 et dont les courbes ROC sont présentées sur la Figure 4.7. Les jeux de données pour l'évaluation niveau pixel contiennent des masques binaires qui mettent en évidence les positions où sont localisés les événements. Pour le jeu Ped 1, les vérités terrains ne sont disponibles que pour une dizaine de vidéos. L'ensemble des vidéos du jeu Ped 2 est accompagné de vérités terrains. Les performances notées sont faibles sur le jeu de données Ped 1 et bonnes sur le jeu de données Ped 2. La concordance entre les résultats de l'évaluation niveau frame et pixel sur la base Ped 2 s'explique par le fait que l'évaluation est faite sur les blocs de mouvements saillants qui couvrent assez bien les événements détectés. Autrement dit, tous les événements détectés sont couverts à plus de 40% par les blocs de mouvements saillants. L'AUC global donne une indication sur la performance de l'approche à détecter les événements

Tableau 4.4 – Evaluation niveau Pixel avec le noyau linéaire

	Ped 1		Ped 2	
Méthodes	**AUC**	**RD**	**AUC**	**RD**
Saillance-Spatiale	0,58	0,56	0,77	0,69
Saillance-SpatioTemporelle	0,58	0,55	0,79	0,70
HOOF-Spatiale	0,56	0,53	0,69	0,63
HOOF-SpatioTemporelle	0,56	0,52	0,75	0,67

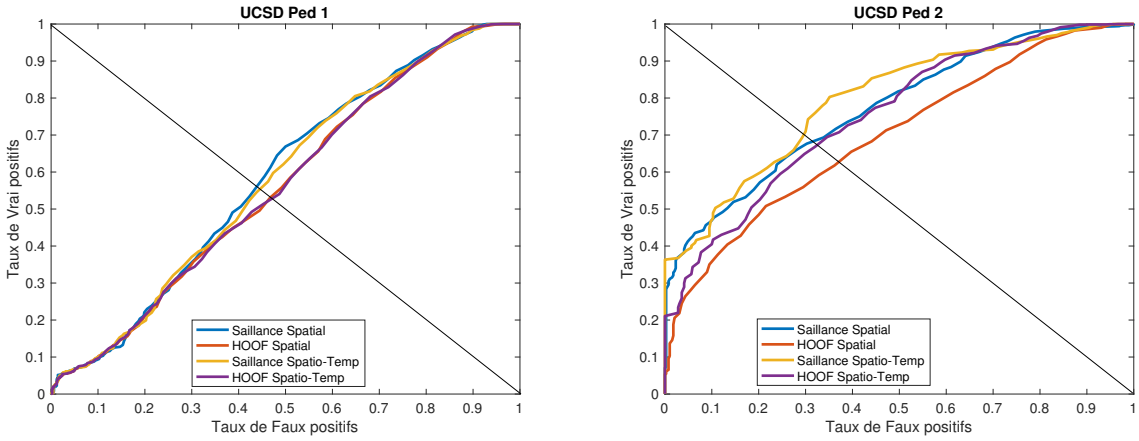


FIGURE 4.8 – Courbes ROC : Evaluation niveau Pixel

rares. Etant donné qu'il peut exister une grande variabilité inter-classe d'événements au niveau des vidéos de la base, une analyse de la performance par vidéo, donne une idée plus claire des événements difficiles à détecter et ainsi élaborer des hypothèses sur les causes probables de l'échec de détection. Ainsi, les tableaux 4.5, 4.6 et 4.7 donnent les UAC calculés individuellement pour chaque vidéo de chaque base. Il ressort de l'analyse de ces résultats plusieurs observations. Sur la base UCSD Ped 1, 24/36 des vidéos de la base ont obtenues un AUC supérieur à 80% pour la saillance contre 21/36 pour le descripteur HOOF. La saillance obtient une moyenne de $92,83\% \pm 6,04$ tandis que le descripteur HOOF obtient une moyenne de $92,84\% \pm 5,78$. Sur la base UCSD Ped 2, les deux méthodes présentent des résultats satisfaisants sur la plupart des vidéos. Il faut noter qu'il n'est pas possible de calculer l'AUC individuellement pour les vidéos 8, 9, 10 et 11 car elles ne contiennent que des séquences d'images d'événements anormaux. En effet, le calcul de l'AUC nécessite de disposer de données de test avec au moins deux classes. La visualisation des vidéos ayant obtenu des taux de détection inférieurs à 80% nous a permis de donner dans le tableau 4.8 quelques commentaires sur les événements difficiles à détecter.

Tableau 4.5 – Evaluation niveau frame de l'approche : Base UCSD Ped 1

	AUC																	
Vidéos	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
Saillance	0,96	0,90	0,69	0,39	0,63	0,76	0,90	0,97	0,90	0,98	0,75	0,76	0,96	0,94	0,80	0,85	0,70	0,41
HOOF	0,95	0,90	0,69	0,40	0,62	0,69	0,76	0,96	0,92	0,99	0,63	0,79	0,93	0,87	0,84	0,78	0,83	0,37

Tableau 4.6 – Evaluation niveau frame de l'approche : Base UCSD Ped 1 Suite

	AUC																	
Vidéos	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33	V34	V35	V36
Saillance	0,98	0,68	0,91	0,96	0,25	0,86	0,99	0,18	0,95	0,95	0,94	0,89	0,99	0,98	0,98	0,80	0,72	0,81
HOOF	0,88	0,64	0,79	0,94	0,28	0,89	0,98	0,24	0,88	0,97	0,97	0,94	1	0,97	0,99	0,80	0,74	0,72

Tableau 4.7 – Evaluation niveau frame de l'approche : Base UCSD Ped 2

	AUC											
Vidéos	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
Saillance	0,70	1	0,51	0,99	0,99	0,79	0,34	-	-	-	-	0,23
HOOF	0,87	1	0,35	0,98	0,99	0,64	0,67	-	-	-	-	0,45

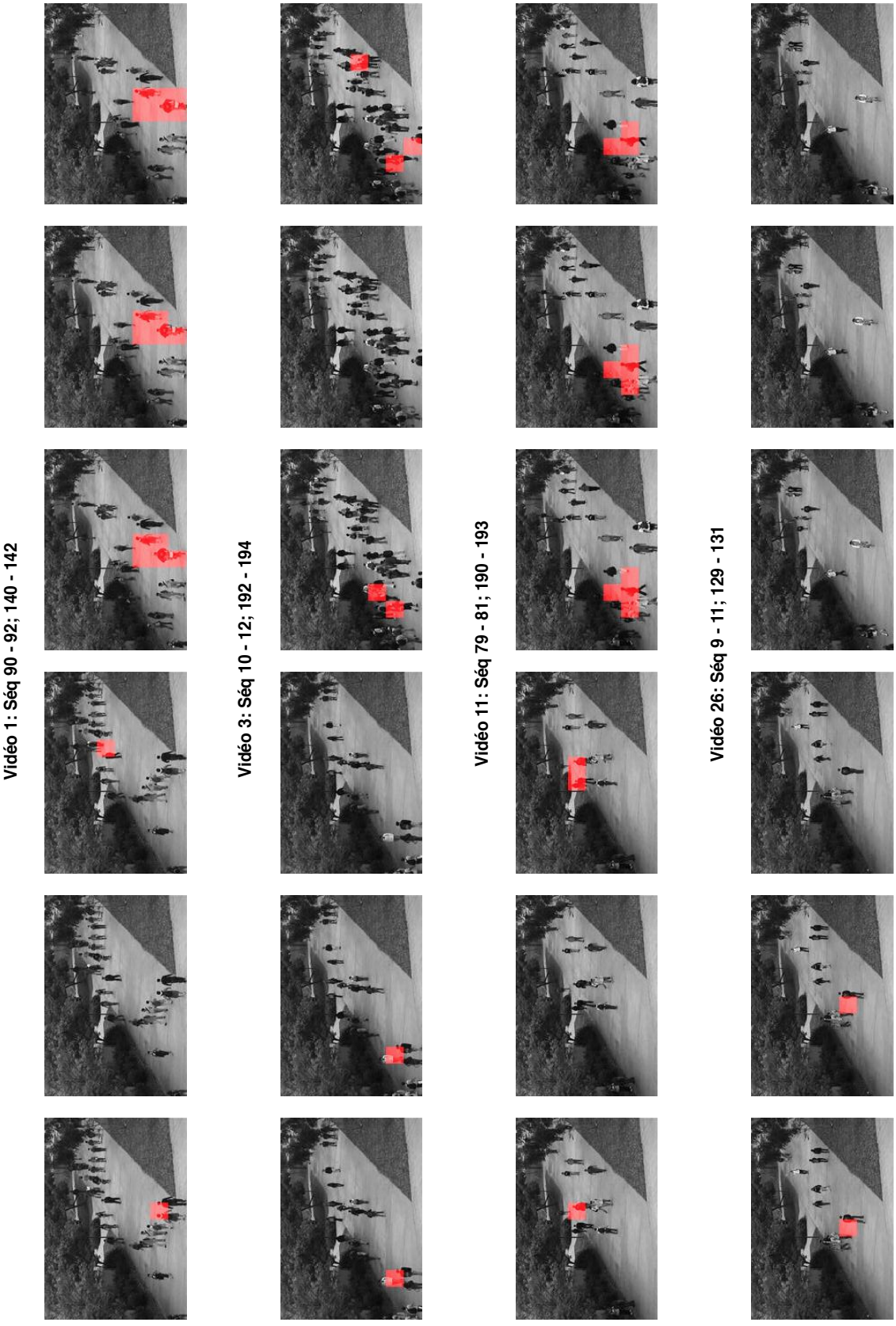


FIGURE 4.9 – Exemple de localisation d'Evénements sur la base Ped 1

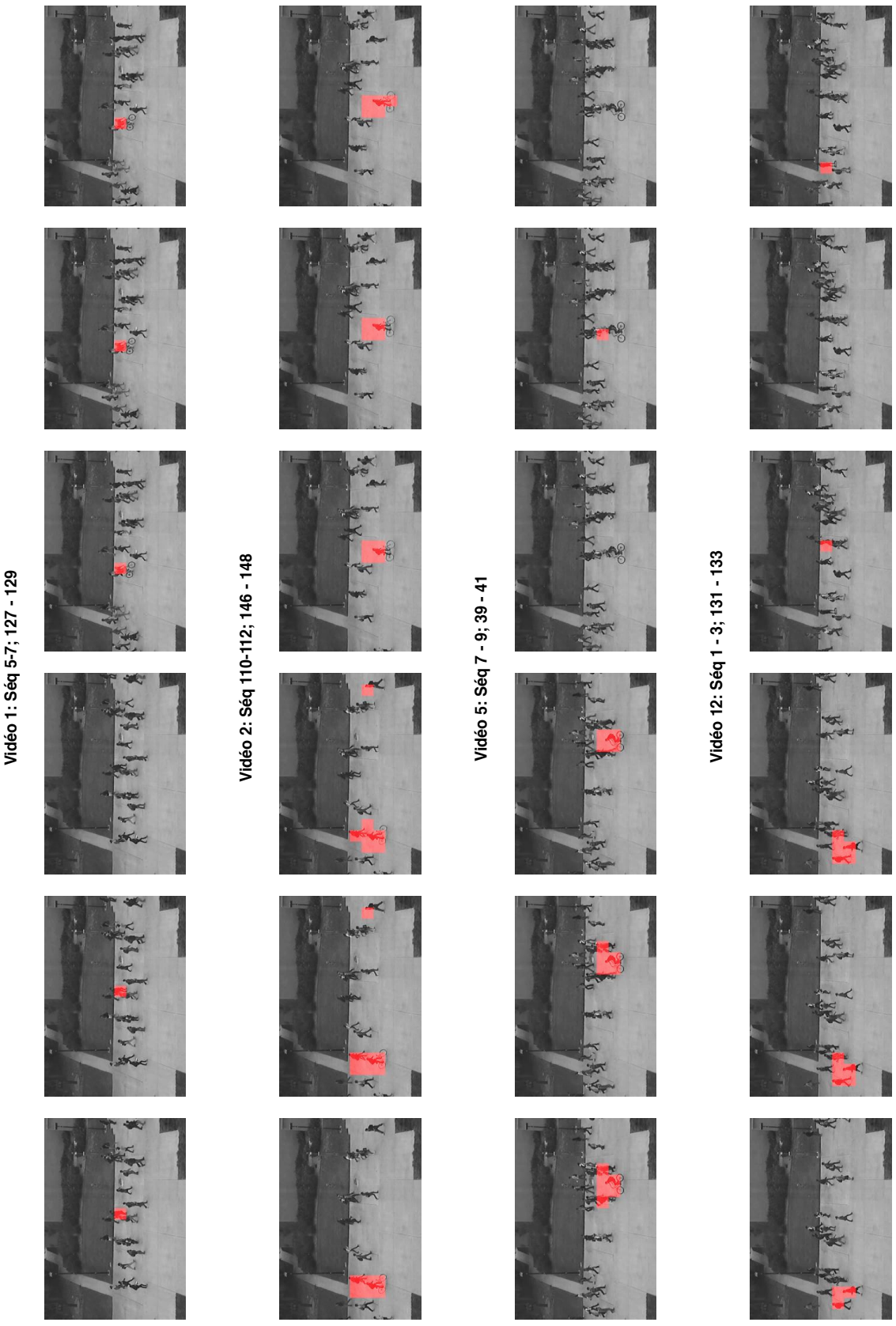


FIGURE 4.10 – Exemple de localisation d'Evénements sur la base Ped 2

Nous constatons de ces observations que la plupart des événements non détectés sont liés à la vitesse des entités qui les génèrent. Un individu sur une trottinette possède dans bon nombre de cas une faible vitesse et est facilement confondable avec un piéton. D'autres groupes d'événements non détectés ont en commun la forte densité de piétons dans la scène. Ces observations sont également valables pour la deuxième base UCSD Ped2 pour les quelques vidéos sur lesquelles l'approche a donné de faible score d'AUC.

Tableau 4.8 – Observations sur quelques vidéos de la base Ped 1

Vidéos	Observations
4, 18	Anomalie liée au passage d'une personne sur une trottinette
23	Scène presque vide avec le passage d'une personne en fauteuil électrique
26	Scène presque vide avec passage d'un cycliste

Les figures 4.9 et 4.10 montrent quelques séquences d'images de localisation des événements effectuées à partir du descripteur de saillance. Dans ces séquences d'images, les emplacements des événements rares sont marqués au rouge. Il s'agit des blocs ayant obtenus un score supérieur au seuil optimal déterminé à partir de la courbe ROC. Il faut noter que ces cartes de localisations sont proches des cartes de saillance à la différence qu'il s'agit ici de déterminer les blocs qui contiennent des événements rares parmi ceux contenant des mouvements saillants. Ces quelques échantillons de résultats montrent que l'approche arrive à localiser les événements mais dans certains cas se trompe. Ces échantillons de résultats permettent de mettre en évidence certains cas où l'approche se trompe sur la localisation précise de l'anormalité. Les séquences $131-134$ de la vidéo 12 de la base Ped 2 illustrent parfaitement l'un de ces cas. Les séquences $129-131$ de la vidéo 26 de la base Ped 1 illustrent un cas de non détection et de non localisation de l'événement.

B - Etude de l'influence du noyau du SVM

Dans la sous-section précédente, nous avons effectué une modélisation en utilisant un noyau linéaire du SVM one class. Le but de cette sous-section est d'observer l'influence du noyau de projection de données du SVM sur les performances de l'approche. Ainsi, nous avons utilisé le noyau gaussien et le noyau polynomial pour la construction des modèles. Les résultats obtenus sont présentés sur la figure 4.11. Ils montrent pour chaque noyau utilisé, les valeurs de l'AUC pour une évaluation niveau frame sur chacune des bases Ped. Il ressort de ces résultats que le noyau n'a pas particulièrement d'effets sur les performances de l'approche. Les valeurs des AUC sont sensiblement égales. Autrement dit, quel que soit le noyau utilisé pour la construction du modèle, les performances sont sensiblement les mêmes. Outre le critère de performance, celui du temps de calcul peut peser dans la balance. Sur ce dernier point, le noyau linéaire est le plus rapide des trois.

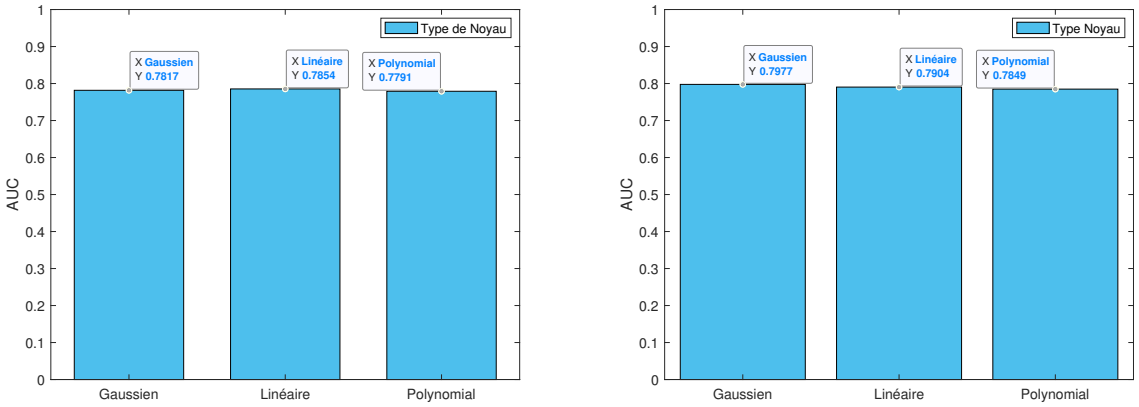


FIGURE 4.11 – Influence du noyau SVM (à gauche Ped1 et à droide Ped 2)

C - Incorporation des informations de formes

Au vu des types d'événements de la base UCSD, il est remarquable que la plupart des événements rares sont liés à la vitesse des entités présentes. Néanmoins, certains sont dus à l'apparence de ces entités dans les vidéos. Cette sous-section a pour objectif d'étudier l'apport de certains descripteurs de formes étudiés dans les chapitres précédents. Afin de pouvoir fusionner les informations de mouvement et d'apparence, nous avons opté pour deux schémas de fusion : le premier concerne la concaténation simple entre les vecteurs de saillance et les différents vecteurs de descripteurs de forme et le second est une fusion au niveau des scores de décision issus de l'algorithme de prédiction. Les paramètres des différents descripteurs utilisés sont détaillés :

- la taille des blocs est fixée à 20×20 :

- **HOG** : la taille de l'histogramme est de 32, obtenue à partir des différentes valeurs des paramètres liées au descripteur ;

- **CHOP** : la taille du vecteur est de 36 ;

- **LBP** : les caractéristiques de LBP sont extraites en considérant des 8 pixels pour le voisinages.

- **Noyau SVM** : Pour cette expérimentation, nous avons utilisé un noyau linéaire.

Les différents résultats obtenus sont consignés dans le tableau 4.9

Tableau 4.9 – Evaluation niveau Frame : Combinaison de descripteurs

Type de fusion	**Concaténation**				**Fusion de score**			
Base de données	**Ped 1**		**Ped 2**		**Ped 1**		**Ped 2**	
Critères / **Méthodes**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**
Saillance	0,78	0,28	0,79	0,30	0,7854	0,2817	0,7904	0,3021
Saillance-HOG	0,60	0,43	0,64	0,38	0,7812	0,2896	0,7941	0,3122
Saillance-CHOP	0,58	0,45	0,63	0,40	0,7641	0,3023	0,7883	0,2948
Saillance-LBP	0,60	0,43	0,66	0,35	0,7826	0,2903	0,7907	0,2945

Contrairement à ce que nous pouvions espérer de la fusion des informations de mouvement avec celles de forme, il n'y a pas eu d'amélioration majeure au niveau des performances. Pire, pour certains descripteurs, leur incoporation dans le modèle a provoqué une baisse de performance. Néanmoins, on peut tirer certaines informations de ces différentes fusions. En effet, la concaténation des vecteurs de descripteurs dégrade énormément les performances de l'approche. Ce type de fusion de notre descripteur de saillance avec les différents descripteurs de forme est à éviter. Pour ce qui est de la fusion au niveau des scores de décision, elle est prometteuse. Elle permet par endroit d'améliorer les performances. Une plus grande attention pourra être apportée à ce type de fusion pour éventuellement tirer le meilleur des deux types d'informations.

D - Comparaison avec l'état de l'art

De nombreuses méthodes ont été proposées dans l'état de l'art pour améliorer les performances de détection. Deux grands groupes de travaux peuvent être identifiés : le groupe des approches classiques auquel appartient notre approche et récemment le groupe des approches basées sur l'apprentissage profond. Le tableau 4.10 présente le résumé de différentes méthodes de la première catégorie de la littérature pour une comparaison avec notre approche. Il faut souligner qu'une comparaison objective des différentes méthodes est difficile compte tenu de plusieurs facteurs à savoir :

- La non disponibilité dans la plupart des cas des codes des approches pour exécution ;
- La divergence d'un article à un autre des résultats pour des approches ayant été proposées avant la mise à disposition de la plupart des bases de données ;
- Les problèmes de labélisation au niveau de certaines bases de données :
- La multiplicité des critères d'évaluation.

Tableau 4.10 – Comparaison avec les méthodes existantes

Base de données	**Ped 1**				**Ped2**			
Niveau	**Frame**		**Pixel**		**Frame**		**Pixel**	
Critères / **Méthodes**	**AUC**	**EER**	**AUC**	**RD**	**AUC**	**EER**	**AUC**	**RD**
MDT temporel [81]	0,82	0,23	0,57	0,59	0,76	0,28	0,52	0,57
MDT spatial [81]	0,6	0,43	0,66	0,5	0,75	0,29	0,66	0,63
Social Force [81]	0,69	0,36	0,22	0,41	0,7	0,35	0,22	0,28
MPPCA [81]	0,67	0,35	0,22	0,23	0,71	0,36	0,22	0,22
Adam (MHL) [81]	0,63	0,38	0,16	0,42	0,58	0,46	0,16	0,22
Energy-based [121]	0,7	0,35	0,49	0,67	0,86	0,16	0,72	0,84
Zhang et al. [122]	0,86	0,19	0,76	0,74	-	-	-	-
SA-MHOF [123]	0,86	0,19	0,63	0,68	-	-	-	-
TCP [115]	0,96	0,08	0,64	0,41	0,88	0,18	-	-
Amraee et al. [124]	0,85	-	0,68	-	0,93	-	0,85	-
Notre Approche	0,78	0,28	0,58	0,56	0,8	0,3	0,79	0,69

Dans le tableau 4.10, les tirets ("−") représentent les données manquantes, c'est à dire non fournies par les auteurs. On peut noter que les performances de notre approche se situent dans la moyenne des performances réalisées sur la base UCSD. Nos performances dépassent celles consignées pour les méthodes de Mixture Dynamique de Texture (MDT) spatial, de force sociale, MPPCA, Adam. Une comparaison du temps de calcul est quasi impossible compte tenu des difficultés énumérées ci-dessus.

4.4 Conclusion

Le chapitre a été consacré à la détection et à la localisation des événements rares dans une scène. La démarche que nous avons adoptée a consisté à la mise en place d'un algorithme de mise en évidence de mouvements saillants dans une vidéo. Cet algorithme exploite les propriétés de la transformée en cosinus discrète (TCD) pour déterminer les mouvements irréguliers dans une scène au regard de tous les mouvements qui s'y déroulent. L'algorithme a montré de bonnes performances pour la mise en évidence des mouvements saillants. Par la suite, nous avons proposé une approche complète de modélisation locale d'évènements dans le but de détecter et de localiser ceux qui sont rares ou anormaux. Cette approche exploite les capacités du précédent algorithme pour d'une part restreindre les zones de recherche dans une scène en se focalisant uniquement sur les zones contenant des mouvements saillants et d'autre part en utilisant les scores de saillance non normalisés comme descripteurs pour la construction de modèle d'événements. Dans cette dernière étape, nous avons eu recours à l'algorithme "SVM one class" qui permet grâce aux fonctions noyau de définir une frontière autour des jeux de données d'apprentissage.

Nous avons procédé à différentes évaluations de notre approche sur des bases de données publiques abondamment utilisées dans la littérature. Les performances de notre approche sont équivalentes à celles de l'état de l'art. Le haut du tableau des performances est occupé par des méthodes plus complexes. Nous avons également présenté des résultats issus de tentative d'amélioration de performances qui peuvent servir dans la littérature comme guide pour de nouveaux travaux. En effet, l'incorporation des informations de forme pour améliorer la détection notamment celle des événements qui seraient liés à la forme des entités présentes dans la scène. La fusion par concaténation n'a pas été probante car cela réduit énormement les performances de l'approche. A l'opposé, la fusion dite "tardive" qui se fait au niveau des scores de décision est prometteuse car elle ne dégrade pas les performances mais les améliore par endroit. Une étude plus profonde peut être menée pour voir comment améliorer l'apport de ces informations de forme dans la détection.

Pour finir, nous avons présenté quelques résultats qualitatifs pour montrer aussi bien les bonnes localisation que les mauvaises. Une étude des causes d'échec de l'approche sur certaines vidéos a été faite. De cette étude, nous avons pu observer que notre approche est tributaire de la qualité des flots optiques extraits dans la scène. En effet, dans la plupart des cas d'échec, les anormalités sont liés à des entités avec de faible vitesse ou avec des apparences qui se confondent facilement avec le décor. Le choix d'un meilleur algorithme de flot optique est important quand à l'application de notre approche. L'avantage de notre approche est qu'elle peut être adaptée à d'autres type d'informations comme le gradient par exemple. Ainsi, une étude de l'adaptation de l'approche à d'autres caractéristiques pourrait peut-être apporter une solution à ce problème.

Conclusions et Perspectives

Le travail présenté dans ce rapport porte sur le développement d'approches complètes pour la détection et la localisation des événements rares dans une vidéo. Ce travail a donné lieu à trois différentes approches et à diverses études. Nous donnons dans ce chapitre le récapitulatif des différentes contributions suivies des perspectives de travaux futurs sur la détection, la localisation et l'analyse des événements rares.

Travaux réalisés

Etude de descripteurs de forme

La détection d'objets ou de personnes dans une scène passe d'une part par l'extraction de caractéristiques pertinentes et d'autre part par la modélisation de ces caractéristiques. Il en est de même pour la détection et la localisation d'événements rares dans une scène. La qualité des caractéristiques extraites inflence directement sur les performances des algorithmes de détection. Dans cette thèse, nous avons étudié la robustesse de plusieurs descripteurs utilisés pour la détection de personnes, d'objets, etc. Deux descripteurs ont attiré notre attention : l'histogramme des orientations du gradient et l'histogramme des orientations de phase. Une partie des travaux est consacrée à l'étude de la robustesse de ces deux descripteurs face à la contrainte de variation des conditions de luminosité. Nous avons mené cette étude pour déterminer les descripteurs les mieux adaptés au cadre applicatif de la thèse qu'est la détection dans des conditions de luminosité changeantes. Il ressort de notre étude que l'histogramme des orientations des phases est plus robuste aux variations de la luminosité. Ces descripteurs ont servis dans les travaux consacrés à l'élaboration d'approches de modélisation des événements globaux ou locaux.

Détection d'événements globaux

Les travaux sur la détection d'événements rares sont scindés en deux catégories. La modélisation globale de la scène pour la détection des événements globaux constitue la première catégorie d'approches. Par événements globaux, nous sous-entendons les événements qui se déroulent simultanément dans toute la scène comme par exemple les mouvements de foule. Le défi d'une telle détection est d'identifier dans une vidéo, les moments exacts de démarrage et de fin des événements. Nous avons proposé une approche baptisée BGMMAI pour "Bayesian Generative Model based on Motion and Appearance Information". Dans cette approche, nous proposons une extraction locale des caractéristiques de la scène suivie d'une modélisation globale. Deux types

de caractéristiques sont extraites de la scène : les caractéristiques de mouvements et d'apparence. La modélisation des événements avec les descripteurs retenus est faite en utilisant l'algorithme d'allocation latente de Dirichlet (LDA). L'utilisation de cet algorithme permet de faire émerger de façon non supervisée des groupes sémantiques de mouvements présents dans la base d'apprentissage. Pour y parvenir nous avons employé l'approche "sac-de-mot" pour transformer les vecteurs de descripteurs d'un volume d'images en un document visuel représenté par un vecteur unique. L'identification des volumes d'images suspects se fait en calculant la vraissemblance du document visuel qui le représente par rapport au modèle du corpus (base d'apprentissage) appris. A travers les expériences, nous avons montré que pour notre approche permettait d'améliorer les performances. Nous avons également montré que la meilleure combinaison de descripteurs est celle du descripteur HOOF avec le descripteur CHOP. Une évaluation de l'influence de différents paramètres qui entrent en jeu a été faite. Il ressort de nos études que l'approche est relativement stable et donc ne dépend pas de façon critique d'un paramètre précis.

Localisation de mouvements saillants dans une vidéo

La localisation des événements dans une scène consiste à rechercher dans celle-ci des événements qui, d'une part se détachent du reste des événements en cours dans la scène et, d'autre part, qui ne ressemblent pas à ceux appris. Pour aborder la localisation des événements, nous nous sommes penchés dans un premier temps sur la localisation des mouvements saillants de la scène. Par mouvements saillants, il faut comprendre les mouvements qui se différencient soit par leur vitesse, soit par leur orientation ou soit par les deux simultanément. Ainsi, nous avons proposé une méthode pour automatiquement les mettre en évidence. Cette dernière est entièrement non supervisée car ne nécessitant pas d'étape d'apprentissage. Elle exploite les propriétés de regroupement d'énergie de la transformée en cosinus discrète (TCD). La reconstruction du signal original à partir des signes des coefficients de la TCD permet d'obtenir une carte des irrégularités de la matrice. Nous avons appliqué cette propriété aux données de vitesse et d'orientation du flot optique pour obtenir leur carte d'irrégularité. La combinaison des deux cartes nous permet d'obtenir une carte unique des mouvements irréguliés de la scène. Les résultats présentés dans le chapitre 4 montrent une bonne détection des mouvements saillants de la base UCSD Ped 2.

Détection et Localisation des événements locaux

A la suite de la localisation des mouvements saillants, nous avons montré que les résultats obtenus peuvent servir dans une approche de localisation d'événements. Dans l'approche que nous avons proposé pour la détection et la localisation des événements rares, nous avons remplacer l'algorithme de filtrage des points d'intérêt de la méthode BGMMAI par la méthode de détection des mouvements saillants. A la place des points d'intérêts, des blocs ou volumes de blocs sont extraits et filtrés grâce à leur score de saillance qui traduit la présence ou non de mouvements saillants au sein de ces derniers. Les différentes caractéristiques sont extraites à l'intérieur de ces blocs. Contrairement à la modélisation globale proposée pour la méthode BGMMAI, nous avons utilisé la version "one class" de l'algorithme SVM pour modéliser les caractéristiques. L'utilisation

de cet algorithme permet de définir une frontière autour des vecteurs de caractéristiques des événements normaux contenus dans la base d'apprentissage. Au niveau des caractéristiques, nous avons démontré que l'utilisation des scores de saillance des pixels des blocs donne des résultats pertinents et comparables à l'histogramme des orientations du flot optique. Nous avons également étudié l'apport des descripteurs de formes dans notre approche. Les résultats montrent que certains de ces descripteurs n'apportent qu'une légère amélioration à la performance de l'approche tandis que d'autres dégradent cette performance. De l'analyse de ces résultats, il ressort que l'intégration des informations d'apparence telle que réalisée n'est pas pertinente.

Intégration des algorithmes dans un démonstrateur

Les algorithmes proposés durant nos travaux ont été intégrés dans un système de gestion automatique de l'éclairage public et y ont apporté une plus value par rapport aux systèmes exitants, car intégrant une analyse complète de la scène avec détection des événements rares. En effet, les systèmes intelligents actuellement proposés sur le marché ou en cours d'étude sont principalement basés sur les détecteurs de mouvement (comme les capteurs infrarouges passifs ou les détecteurs pyroélectriques). Nous pouvons, par exemple citer le projet SMARTLIGHT supporté par l'ADEME. Porté par un industriel, l'objectif du projet est d'équiper les luminaires d'un détecteur de présence, d'un capteur de lumière du jour, d'un outil de transmission sans fil et d'une intelligence embarquée. Le système GEPPADI, testé en Belgique, permet de détecter la présence d'un usager, d'en identifier le type en fonction de sa vitesse (piéton, cycliste, jogger, véhicule) et de communiquer avec les autres lampadaires pour éclairer petit à petit la route. Plus récemment, la société Philips a mis au point le système LumiMotion basé sur l'utilisation d'une caméra 360°. Ce système permet la détection de piétons, de cyclistes et de voitures lentes. Il est programmé pour différents scénarios. À San Diego, le système d'éclairage urbain utilise une gradation de l'intensité lumineuse en fonction de l'horaire (plus faible en pleine nuit) et des phases critiques (levé et couché du soleil).

Perspectives

Base de données réaliste

Les différentes bases de données utilisées dans cette thèse pour l'évaluation des différentes méthodes ne sont pas représentative de la réalité. La détection d'événements rares dans un contexte urbain ou péri-urbain présente plus de challenges que ceux rencontrés sur des bases disponibles et exploitées dans cette thèse. Les conditions d'acquisition des données dans un cas réel tel que la surveillance routière par exemple, sont plus complexes en ce sens qu'il faut prendre en compte un certains nombre de contraintes supplémentaires. Ces contraintes peuvent être liées aux conditions météorologiques qui varient selon la période de l'année et même de la journée où la détection est effectuée. Il y a également la densité de la circulation, la vitesse des usagés de la route, la variabilité des événements rares qui peuvent s'ajouter aux contraintes sus-citées. Outre ces contraintes liées à l'environnement, des contraintes techniques comme la qualité et la taille

des images fournies par les caméras doivent être prises en compte pour une détection temps réel. Il serait alors intéressant de créer une base de données qui intégre ces différentes contraintes pour une évaluation plus réaliste des différentes approches proposées et celles de l'état de l'art.

Qualité des descripteurs

La qualité des descripteurs impacte directement les performances des méthodes de détection. Dans cette thèse, nous avons présenté un certains nombre de descripteurs de mouvement et d'apparence. Les descripteurs de mouvement que nous avons utilisés sont basés sur le flot optique comme d'ailleurs beaucoup d'autres descripteurs dans la littérature. De nombreux travaux sur l'extraction du flot optique existent dans la littérature avec une amélioration continuelle des performances face à des conditions de détection variées. Malgré ces bonnes performances, l'exploitation d'informations différentes de celles du flot optique peut s'avérer payante. Une étude sur l'impact du remplacement du flot optique par d'autres informations pourra être envisagée. Quant aux descripteurs de formes, nous nous sommes limités aux descripteurs traditionnels. Les descripteurs tels que ceux issus des réseaux de neurones peuvent permettre une amélioration des performances. Dans un premier temps, l'évaluation de la robustesse des descripteurs face aux différentes conditions de luminosité peut être étendue à ces descripteurs. Dans un second, leur incorporation dans les approches de détection d'événements pourra être étudiée.

Amélioration de la détection de mouvements saillants

Une partie de cette thèse a été consacrée à la localisation de mouvements saillants dans une scène. L'approche que nous avons proposé se base sur la transformée en cosinus discrète. Dans cette approche, le calcul du flot optique se fait à une seule échelle et par conséquent, l'analyse des mouvements également. Une extraction et une analyse en pyramide du flot optique serait intéressant pour la prise en compte de petits mouvements difficilement détectables. Ainsi, une version multi-échelle de la méthode peut ouvrir la voie à une amélioration des performances de détection.

Il est connu qu'à part les transformées de Fourier et de consinus, il existe bien d'autres tranformées. Il peut être envisagé de changer la TCD dans notre approche par d'autres types de transformée comme par exemple la transformée de Mojette utilisée par Coudert et al. [125] pour des tâches similaires.

Prise en compte des énévements statiques

L'aspect du problème des évènements rares non pris en compte dans cette thèse est celui des évènements générés par des entités statiques. Nous avons fait l'hypothèse que les événements sont générés par des entités avec des mouvements saillants. Pour intégrer la détection des événements liés à des entités statiques, l'utilisation d'algorithmes de soustraction d'arrière-plan peut s'avérer importante pour d'abord la détection de ces entités en vue de leur traitement. On pourrait envisagé la mise en place de deux modèles complémentaires de détection l'un pour les entités mobiles et le second pour les entités immobiles.

Amélioration de la fusion des descripteurs

La prise en compte des informations d'apparence nous a amené à expérimenter quelques schémas de fusion des caractéristiques. Les résultats présentés dans cette thèse montrent un faible apport des différentes fusions effectuées entre les caractéristiques de mouvement et d'apparence. Il serait intéressant d'étudier d'autres moyens de fusion afin de profiter pleinement des informations qu'apportent les descripteurs de forme.

Vers les approches d'apprentissage profond

L'apprentissage profond ne se limite pas à l'extraction de caractéristiques de haut niveau pour la détection et la classification d'objets. Comme nous l'avons montré dans le chapitre 2, de plus en plus de travaux sur la détection d'événements rares se tournent vers l'utilisation de différents types de réseaux de neurones. Parmi ces différentes approches, celle utilisant le GAN (Generative Adversarial Networks) nous semble prometteuse. Elle permet de contourner certains obstacles tels que le manque de données d'apprentissage, l'inexistance de jeu de données sur les événements rares dans les données d'apprentissage. Ces obstacles rendent difficile l'utilisation des approches basées sur les CNN ou RNN pour ce type de problème. Pour détecter les événements rares, les GAN apprenent à générer de nouvelles données à partir uniquement des données disponibles sur les événements fréquents. Un exemple est la génération de données de flot optique à partir des flots optiques des événements fréquents. Dans un jeu de génération-discrimination, le réseau finit par apprendre à générer et distinguer parfaitement les flots optiques des événéments fréquents, des autres. Le challenge que présente cette approche se trouve au niveau de la recherche du discriminateur optimal qui pourra parfaitement le rôle. Des travaux futurs devront être consacrés à cette problématique. La mise en place de meilleurs réseaux aussi bien pour le générateur que pour le discriminateur impactera les performances de la détection des événements rares.

Annexe A

Etude comparative de moments géométriques : Application à l'estimation de pose humaine 3D

Les moments géométriques sont des descripteurs de formes globaux de type "region-based". Ils ont été abondamment utilisés dans la littérature pour diverses tâches relatives à la description des régions d'images pour la recherche de formes connues, telles que la reconnaissance de caractères, l'analyse d'empruntes digitales, etc. Les moments géométriques modélisent la distribution des pixels dans l'image pour fournir des informations géométriques intéressantes sur la forme analysée. Les moments de petit ordre sont facilement interprétables mais le deviennent moins au fur à mesure que les ordres s'élèvent. Par exemple :

- $m_{0,0}$ représente la masse de l'image et pour des images binaires, représente l'aire de la forme dans l'image ;

- m_{01} et m_{10} représentent les moyennes ;

- $(m_{01}/m_{00}, m_{10}/m_{00})$ est la coordonnée du centre de gravité de l'image.

Un tel descripteur global peut être également utile dans la détection d'événements rares. Il présente l'avantage d'être très robuste, peu sensible au bruit. Il préserve l'information et est rapide à calculer.

L'application qui servira de cadre de comparaison des différents moments retenus est relative à l'estimation de pose humaine. L'estimation de pose humaine 3D à partir d'images monoculaires de sa silhouette 2D, vise à donner une valeur approximative dans le repère 3D des joints qui forment son squelette, en n'utilisant qu'une vue monoculaire de sa silhouette. La méthode que nous avons utilisée dans cette section est issue d'un précédent travail (hors du cadre de cette thèse) et est basée sur les moments géométriques pour l'analyse de la forme des silhouettes. Nous présenterons brièvement la méthode puis ferons une présentation détaillée des moments géométriques dont nous voulons comparer la robustesse dans l'analyse de forme. Le but de cette section n'est donc pas de proposer une méthode d'estimation de pose 3D, mais plutôt une évaluation des moments qui serviront plus tard comme descripteurs en analyse de forme.

A.1 Les moments de Hu

Le premier moment géométrique fut introduit en "vision par ordinateur" par Hu[126]. La formulation générale du moment de Hu d'ordre quelconque $(p+q)$ d'une image 2D est définie comme suit :

$$M_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{A.1}$$

où p et q sont des entiers : $p, q \in \{0, 1, 2, ..., N\}$ et N l'ordre maximal.
Pour rendre le moment invariant à la translation dans le plan image, Hu a démontré que le moment central, faisant intervenir le centre de gravité de l'image, peut être utilisé. L'expression du moment central est :

$$\eta_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{A.2}$$

où

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}}$$

Pour des applications de reconnaissance de forme, Hu a introduit sept moments invariants basés sur les moments centraux. Les six premiers moments encodent la forme avec une invariance à la translation et à la rotation. Le septième moment assure une invariance de biais, que nous pensons, sera utile pour distinguer des formes ayant subi un effet miroir. Les expressions des sept moments se présentent comme suit :

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}$$

Le calcul du moment invariant de Hu est très simple mais présente quelques inconvénients malgré son invariance à la rotation, à l'échelle et à la translation [127] :

- redondance d'information : les moments invariants de Hu présentent un haut dégré de redondance d'information dû à la non orthogonalité de leur base ;
- sensibilité au bruit : les moments de grand ordre sont plus sensibles au bruit ;
- grande variation dans l'intervalle dynamique des valeurs : cette grande variation est due au facteur de puissance dans sa formulation. La conséquence est l'instabilité numérique observable pour des images de grandes dimensions.

Pour une image ou une région d'image donnée, nous calculons les sept moments invariants qui constitueront le vecteur de descripteur de l'image.

$$\mathcal{F}_{Hu} = [\phi_1 \dots \phi_7]^T$$

Pour surmonter les limites associées aux moments géométriques, Teague [128] a suggéré l'utisation de moment orthogonaux continus. Il a introduit deux moments orthogonaux continus : les moments de Zernike et de Legendre, ayant pour bases les polynômes de Zernike et de Legendre respectivement. Des travaux plus récents ont introduit de nouveaux moments orthogonaux pour l'analyse de formes et la reconstruction d'images. Nous présenterons dans la suite trois moments orthogonaux continus.

A.2 Les moments de Zernike

A.2.1 Les Polynômes de Zernike

Largement utilisé en reconnaissance de forme à travers les moments géométriques de Zernike, le polynôme de Zernike forme un jeu complet orthogonal à l'intérieur du cercle unité (voir figA.1). Soit Z_n^m le polynôme de Zernike d'ordre n et de répétition m. Z_n^m est défini par :

$$Z_n^m(\rho, \theta) = R_{nm}(\rho)\exp(jm\theta) \tag{A.3}$$

où

n : un entier positif ;

m : un entier positif ou négatif sous contrainte $n \geq |m|$ et $n - |m|$ pair ;

ρ : Distance radiale normalisée de pixel (x,y) par rapport au centre de la forme ;

θ : Angle d'azimut du pixel (x,y) par rapport au centre de la forme.

Le polynôme radial est défini par :

$$\begin{aligned} R_{mn}(\rho) &= \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s F(n,m,s,r), \\ F(n,m,s,r) &= \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s} \end{aligned} \tag{A.4}$$

$R_{n,-m}(\rho) = R_{n,m}(\rho)$ et tous les polynômes sont soumis à la condition d'orthogonalité :
$\int\int_{x^2+y^2\leq 1}[V_{nm}(x,y)]^* V_{pq}(x,y)dxdy = \frac{\pi}{n+1}.\delta_{np}\delta_{mq}$
avec δ_{ab} la fonction de Kronecher

A.2.2 Moments de Zernike

Les moments 2D de Zernike sont construits à partir du jeu de polynôme combiné avec la fonction d'intensité de l'image. Soit A_{nm} le moment de Zernike d'ordre n et de répétition m. A_{nm} est définie par :

$$A_{nm} = \frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1} f(x,y)V_{nm}^*(\rho,\theta)dxdy \tag{A.5}$$

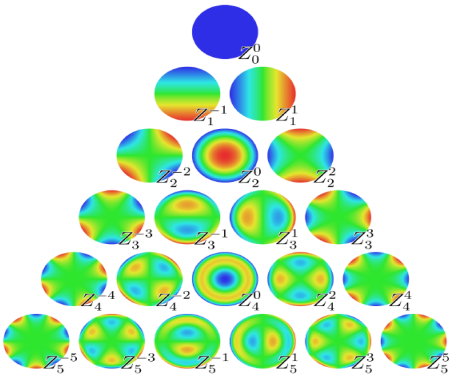


FIGURE A.1 – Polynome de Zernike dans le cercle unité pour différentes valeurs d'ordre et de répétition [129]

Pour une image digitale, l'expression de A_{nm} est donnée ci-après :

$$A_{nm} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f(x,y)[V^*_{nm}(\rho,\theta)] \tag{A.6}$$

où $x^2 + y^2 \leq 1$ et V^*_{nm} est le conjugué du complexe du polynôme.
Les moments de Zernike ont l'avantage d'être robuste face au bruit et aux variations mineures dans la forme, invariant à la rotation et ont une faible redondance d'information. Cependant, son calcul présente quelques inconvénients tels que la normalisation de l'espace des coordonnées, l'approximation continue des intégrales et une grande complexité de calcul [130].

Pour former le vecteur de descripteurs d'une image à partir de ces moments, nous calculons tous les moments possibles d'ordre inférieur à l'ordre donné. Autrement dit, pour un ordre n, nous calculons les moments d'ordre compris entre 0 et n. Cette manière de procéder, nous permet de prendre en compte le maximum d'information possible discriminante pour caractériser la forme dans l'image. L'un des paramètres essentiels dans le calcul des moments est l'ordre maximum convenable pour bien caractériser certaines formes. Plusieurs travaux antérieurs ont étudié les ordres convenables pour une meilleure caractérisation. On peut citer les travaux de Khotanzad et al. [131] qui ont démontré que les ordres convenables se situent entre 7 à 12. Cette étude se base sur les erreurs de reconstruction des images pour plusieurs valeurs d'ordre.

A.3 Les moments orthogonaux de Krawtchouk

A.3.1 Les polynômes de Krawtchouk

Les moments de Krawtchouk ont été introduits en "vision par ordinateur" par P.T Yap et al. [132]. Ils sont calculés en se basant sur les polynômes discrets de Krawtchouk. Les polynômes quant à eux sont basés sur les fonctions hypergéométriques et sont définis comme ci-après :

$$K_n(x;p,N) = \sum_{k=0}^{N} \left(a_{k,n,p} x^k\right) = {}_2F_1\left(-n,-x;-N;\frac{1}{p}\right) \tag{A.7}$$

où $x, n = 0,1,2,...,N \, et \, N > 0, p \in (0,1)$ et la fonction hypergéomtrique est définie comme suit :

$$ {}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \left(\frac{(a)_k (b)_k z^k}{(c)_k} \frac{z^k}{k!} \right) \tag{A.8} $$

$$ (a)_k = a(a+1)...(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{A.9} $$

L'équation (A.9) est le symbole de Pochhammer. L'ensemble de (N+1) polynômes de Krawtchouk forme un jeu complet de fonctions à bases discrètes avec les fonctions de poids :

$$ w(x;p,N) = \left(\begin{array}{c} N \\ x \end{array} \right) p^x (1-p)^{N-x} \tag{A.10} $$

qui satisfait à la condition d'orthogonalité :

$$ \sum_{x=0}^{N} w(x;p,N) K_n(x;p,N) K_m(x;p,N) = \rho(n;p,N)\delta_{nm} \tag{A.11} $$

où $\rho(n;p,N) = (-1)^n \left(\frac{1-p}{p}\right)^n \frac{n!}{(-N)^n}$ et δ_{nm} est la fonction de Kronecher avec :

$$ \delta_{nm} = \left\{ \begin{array}{ll} 1 & si\ n = m \\ 0 & sinon \end{array} \right. $$

Dans le but d'éliminer la grande variabilité de la plage dynamique, un processus de normalisation est appliqué aux polynômes. Ainsi, Yap et al.[132] ont proposé un ensemble de polynômes pondérés normalisés de Krawtchouk défini par :

$$ \bar{K}_n(x;p,N) = K_n(x;p,N) \sqrt{\frac{w(x;p,N)}{\rho(n;p,N)}} \tag{A.12} $$

A.3.2 Les moments de Krawtchouk

En se basant sur les polynômes pondérés, le moment de Krawtchouk d'ordre (n+m) d'une image f de taille N x M est défini par :

$$ Q_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{K}_n(x;p_1,N-1)\, \bar{K}_m(y;p_2,M-1)\, f(x,y) \tag{A.13} $$

Ces moments présentent des avantages que Yap et al. [132] ont mis en évidence dans leur étude. Selon les auteurs, "Les polynômes pondérés d'ordre faible ont des composants qui portent les hautes fréquences spatiales dans une image". Cette propriété combinée avec le fait que les polynômes sont des polynômes à variables discrètes, permet aux moments de Krawtchouk de pouvoir représenter les contours plus efficacement. Les paramètres p_1 et p_2, peuvent être vus comme des facteurs de translation. Ils permettent l'extraction d'information locale dans l'image en permettant un positionnement exacte dans la zone voulue. Cette capacité rend les moments de Krawtchouk semi-locaux, en ce sens qu'ils ne sont pas uniquement extraits sur l'image entière. Pour une forme comme la silhouette humaine, en jouant avec ce paramètre, il est possible de diviser la forme en des parties à analyser individuellement pour améliorer l'information extraite. En effet, si $p = 0.5 + \Delta p$, les polynômes sont déplacés de $N\Delta p$. La direction du déplacement est donnée par le signe de Δp.

Pour extraire les caractéristiques avec les moments de krawtchouk, l'image contenant la forme recherchée est projectée dans les bases polynômiales et les moments sont extraits pour différents ordre. Pour l'application d'estimation de pose, l'analyse de la silhouette humaine est faite en décomposant celle-ci en deux parties (le haut et le bas du corps). Cela a été possible justement en jouant sur les paramètres p_1 et p_2. Pour l'image de la fig. A.2, les valeurs des paramètres sont $p1 = 0.5,\ p2 = 0.1$ (pour le haut) et $p1 = 0.5,\ p2 = 0.95$ (pour le bas). En procédant ainsi, le vecteur de descripteurs final est obtenu par concaténation des vecteurs de caractéristiques des deux parties.

$$\mathcal{F}_{Kr} = \left[Q_{nm}^{bas}, Q_{nm}^{haut}\right]^T$$

avec $m \in [0 : M]$ et $n \in [0 : N]$

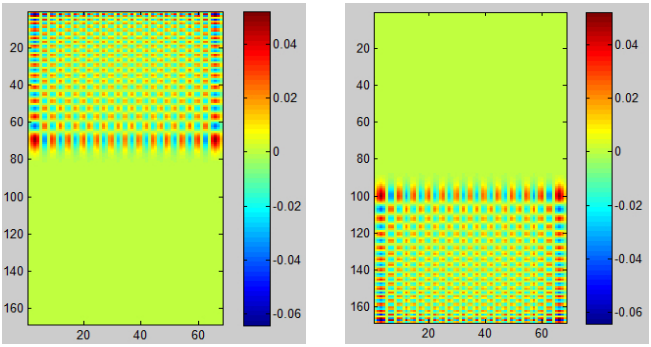


FIGURE A.2 – Polynômes de Krawtchouk pour les moitiés haut et le bas de l'image

A.4 Les moments de Hahn

A.4.1 Polynôme de Hahn

Les moments de Hahn sont une généralisation des moments de Krawtchouk et de Chebyshev [133]. Cela veut dire que les moments de Hahn encapsulent les propriétés des deux autres moments. En utilisant les moments de Hahn pour l'étude comparative, nous cherchons à montrer si ces derniers seront meilleurs que ceux de Krawtchouk dans une application de mise en correspondance de forme (shape matching). Si notre étude prouve une supériorité, ne serait ce que petite, de ce dernier sur celui de Krawtchouk, cela serait une confirmation de la théorie de généralisation citée plus haut. Les moments de Hahn sont également calculés à partir de polynômes basés sur les fonctions hypergéométriques. L'expression des polynômes de Hahn est donnée par l'équation A.14 :

$$h_n(x;\alpha,\beta,N) = {}_3F_2\left(-n, n+\alpha+\beta+1, -x; \alpha+1, -N; 1\right) \tag{A.14}$$

où $\alpha > -1$ & $\beta > -1$

L'ensemble des (N+1) polynômes de Hahn forme un jeu complet de fonctions discrètes pondérées :

$$w(x;\alpha,\beta,N) = \begin{pmatrix} \alpha+x \\ x \end{pmatrix} \begin{pmatrix} \beta+N-x \\ N-x \end{pmatrix} \tag{A.15}$$

et satisfait à la condition d'orthogonalité :

$$\sum_{x=0}^{N} w(x;\alpha,\beta,N)h_n(x;\alpha,\beta,N)h_m(x;\alpha,\beta,N) = \rho(n;p,N)\delta_{nm} \tag{A.16}$$

où

$$\rho(n;\alpha,\beta,N) = \frac{(-1)^n(n+\alpha+\beta+1)^{N+1}(\beta+1)^n n!}{(2n+\alpha+\beta+1)(\alpha+1)^n(-N)^n N!}$$

et δ_{nm} est la fonction de Kronecher.

A.4.2 Les moments de Hahn

Les moments de Hahn ont la même expression que ceux de Krawtchouk à la différences des polynômes. Ainsi, on retrouve dans l'équation A.17 l'expression pour le calcul des moments d'ordre (n+m) pour une image de taille N x M :

$$M_{nm} = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} \bar{h}_n(x;\alpha 1,\beta 1,N-1)\bar{h}_m(y;\alpha 2,\beta 2,M-1)f\,(x,y) \tag{A.17}$$

Le couple de paramètres (α, β) sert à contrôler la sélection de régions spécifiques dans l'image et par conséquent, comme les moments de Krawtchouk, est un descripteur semi-local de type "region-based". Pour le cas particulier des moments de Hahn, le couple $(0,0)$ le rend global. L'extraction de caractéristiques avec les moments de Hahn suit le même processus que celui décrit pour les moments de Krawtchouk. Dans le cas des images de silhouettes, deux régions ont également été utilisées comme pour le descripteur de Krawtchouk (voir fig A.3). La valeur du couple (α,β) est calculée comme dans l'expression ci-après :

$$\alpha_1 = \frac{x_c}{N}t_1 \text{ et } \beta_1 = (1-\frac{x_c}{N})t_1 \text{ suivant l'axe x}$$
$$\alpha_2 = \frac{y_c}{M}t_2 \text{ et } \beta_2 = (1-\frac{y_c}{M})t_2 \text{ suivant l'axe y}$$

où (x_c, y_c) sont les points centraux des régions sélectionnées et les facteurs t_1 et t_2 définient si le moment est global ou local. Pour $t = 0$ le descripteur est global et plus il croit, plus le descripteur devient local.

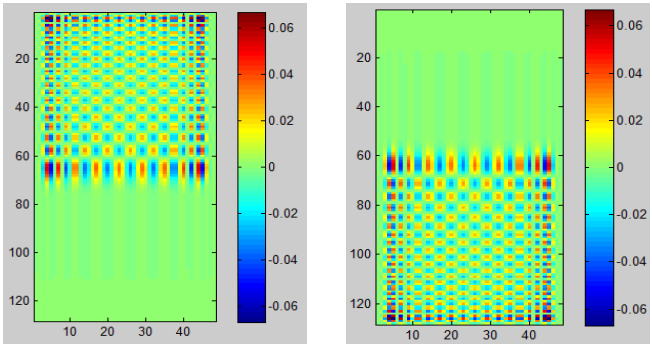


FIGURE A.3 – Polynômes de Hahn pour les moitiés haut et le bas de l'image

Pour notre application d'analyse de silhouette, nous avons déterminé empiriquement que les moments d'ordre 23 donnent de bons résultats. Ainsi, pour une image donnée, on a :

$$\mathcal{F}_{Ha} = \left[M_{0,0}^{bas} \ldots M_{23,23}^{bas}, M_{0,0}^{haut} \ldots M_{23,23}^{haut}\right]^T$$

A.5 Application à l'analyse de silhouettes humaines pour l'estimation de pose 3D

Pour comparer la performance des différents moments présentés ci-dessus, nous allons utiliser une approche simple d'estimation de poste 3D basée sur l'analyse de la forme de silhouettes. Cette méthode se décompose en quatre parties : (1) génération de base de silouhettes et de squelettes simulés, (2) extraction de silhouettes 2D, (3) matching de silhouettes, (4) estimation du squelette 3D et validation.

A.5.1 Méthodologie de l'approche

Comme mentionné plus haut, l'approche est composée de quatre étapes comme le montre également la figure A.4. Les étapes sont détaillées ci-après :

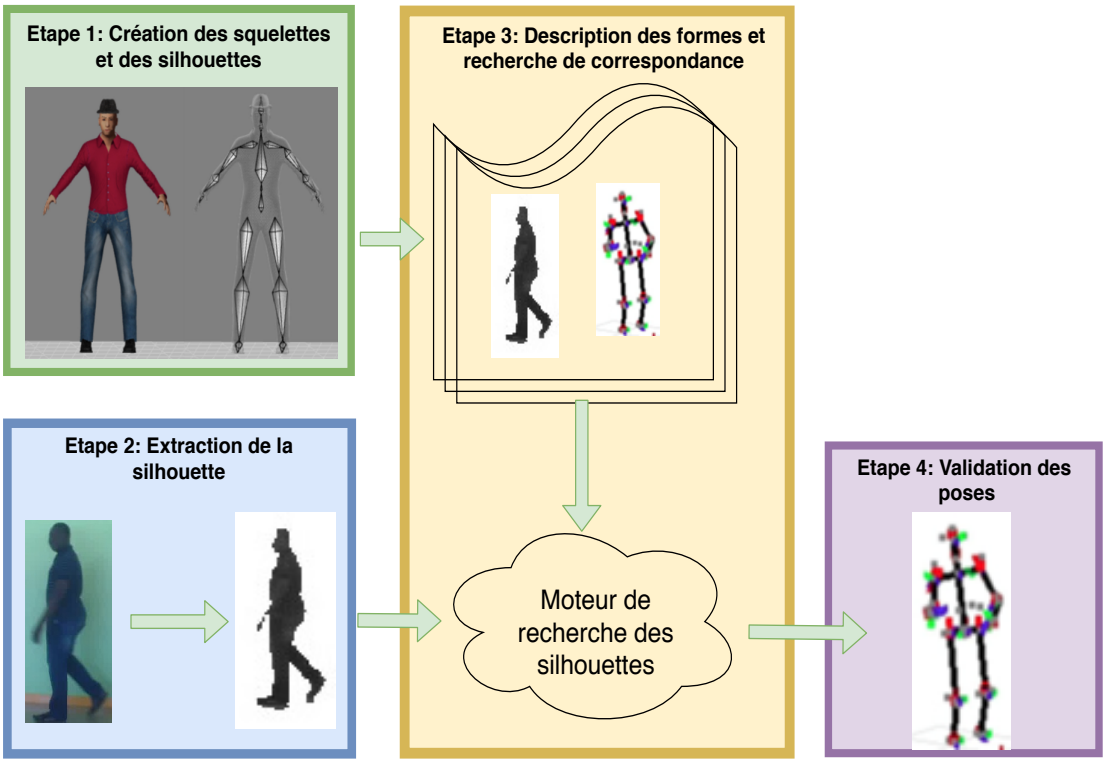


FIGURE A.4 – Méthodologie de l'approche d'estimation de pose 3D

Etape 1 : L'extraction des silhouettes 2D et squelettes 3D pour constituer une base d'apprentissage est réalisée grâce aux logiciels opensource Makehuman [1] (see Fig. A.5), qui sert à générer des modèles virtuels de personnes ayant différentes caractéristiques et Blender [2], afin d'animer les modèles. Les modèles sont animés à partir de données réelles issue de base CMU publique [3]. La base d'apprentissage finale est composée de silhouettes des modèles humains virtuels exécutant des mouvements variés. A chaque silhouette est associée la donnée 3D des poses qui forment le squelette qui le représente.

1. http ://www.makehumancommunity.org/
2. https ://www.blender.org/
3. http ://mocap.cs.cmu.edu/

Etape 2 : La détection de silhouette 2D est assez étudiée dans la littérature. La silhouette est une image binaire qui représente la forme de la personne. Il est préalablement effectué une détection de la personne avant l'extraction de sa silhouette. Ici, il s'agit d'images réelles de personnes dans une position donnée. Les méthodes de soustraction de fond peuvent être appliquées ainsi que les méthodes utilisant des descripeurs présentés plus haut. Les silhouettes sont remises à l'échelle de 48 x 128 pour résoudre les problèmes d'échelle et de translation.

Etape 3 : C'est à cette étape qu'interviennent nos moments pour décrire la forme de la silhouette. Les silhouettes sont finalement représentées par le vecteur de descripteurs extrait à partir des moments géométriques. Pour une silhouette de test, la recherche de la silhouette la plus proche dans la base d'apprentissage est faite en utilisant une messure de distance Euclidienne donnée par :

$$d(z^r, z^n) = \sum_{i=1}^{T} (z_i^r - z_i^n)^2$$

où z^r et z^n sont respectivement les vecteurs de caractéristiques de la silhouette de test et le vecteur de la n^{ime} silhouette dans la base. La recherche des silhouettes dans la base présente ainsi une complexité $O(n)$.

Etape 4 : La mise à l'échelle et l'évaluation sont les dernières étapes de la méthode. Elles consistent à retenir les n plus pertinents résultats ainsi que les squelettes associés. Pour obtenir le squelette final, une moyenne des articulations des n squelettes peut être calculée. Le résultat final est mis à l'échelle de la silhouette requête à travers des transformations géométriques. Le résultat obtenu est ensuite utilisé pour faire une évaluation quantitative et qualitative.

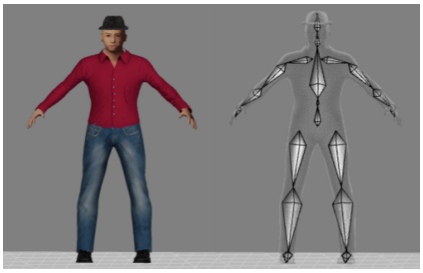


FIGURE A.5 – Modèle virtuel de personne avec le squelette associé

A.5.2 Evaluation

Comme mentionné plus haut, les silhouettes 2D de la base d'apprentissage sont représentées par les vecteurs de descripteurs de moments correspondants ainsi que les squelettes 3D associés. Dans la phase de test, la similarité entre le descripteur de la silhouette requête et ceux de la base d'apprentisage est calculée. Il est possible de ne prendre que le résultat avec une forte similarité (winner-takes-all) ou d'utiliser les k meilleurs résultats pour construire un squelette moyen. L'évaluation de la performance de la méthode est faite en calculant soit une erreur moyenne de reprojection des articulations de la squelette 3D dans l'espace 2D de l'image, soit une erreur moyenne à partir directement des données 3D estimées et connues. Dans le cadre de

cette étude, l'erreur de reprojectoin sera utilisée au détriment de l'erreur obtenue directement des données 3D, étant donner que les vérités terrains 3D ne sont pas toujours disponibles. Un seuil est utilisé pour juger si la pose estimée est correcte. Selon nos analyses, au-delà d'une erreur de 6 pixels, pour les données simulées, l'estimation peut être jugée fausse. Nous rappellons que le but de l'évaluation n'est pas de prouver que la méthode d'estimation est meilleure que celles de la littérature. Il s'agit d'une méthode simple d'estimation de pose pour évaluer la puissance de chaque moment géométrique pour l'encodage des formes. Les poses 3D nous servent à calculer l'erreur qui servira de métrique puisqu'une petite variation de la forme de la silhouette peut engendrer une erreur de reprojection considérable qu'une simple évaluation au niveau de l'image 2D n'aurait pas suffit à détecter. D'autres méthodes telles que le calcul des erreurs de reconstruction d'images, la reconnaissance de caractères, etc. peuvent être utilisées pour les mêmes tâches d'études comparatives.

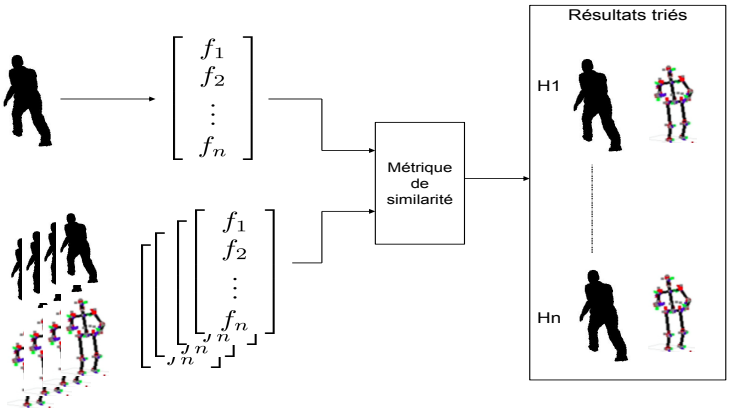


FIGURE A.6 – Processus global de l'expérimentation

La figure A.7 présente un résultat d'une mauvaise estimation obtenu avec le descripteur de Hu. Pour cette silhouette, le descripteur n'a pas été en mesure de bien encoder la forme pour trouver la silhouette similaire dans la base. On peut aussi noter que l'erreur de reprojection correspondante à cette estimation est de 24.84 px qui est largement au-dessus du seuil fixé. Contrairement à ce résultat, la figure A.8 montre des résultats obtenus avec le descripteur Krawtchouk. Ce résultat obtenu est similaire à ceux obtenus à partir de Zernike et Han. Pour cette silhouette, les erreurs de reprojection sont respectivement 0.92 px et 2.69 px.

Le test effectué sur les données simulées de la base montre des résultats intéressants. En testant sur des images non simulées (fig A.9), enrégistrée par nous même à l'aide de smarphone, nous pouvons retrouver des résultats visuellement acceptables dont nous ne pouvons pas calculer l'erreur de reprojection (véritée terrain non disponible). On peut néanmoins noter visuellement que le résultat fourni par le descripteur Hahn semble être plus proche de la pose réelle que les deux autres. Toutefois, sans une évaluation quantitative, la comparaison ne serait pas rigoureuse.

Avant de passer à l'évaluation quantitative, nous avons evalué l'efficacité de l'approche en faisant du suivi de certaines articulations estimées pour une séquence successive d'images d'écrivant

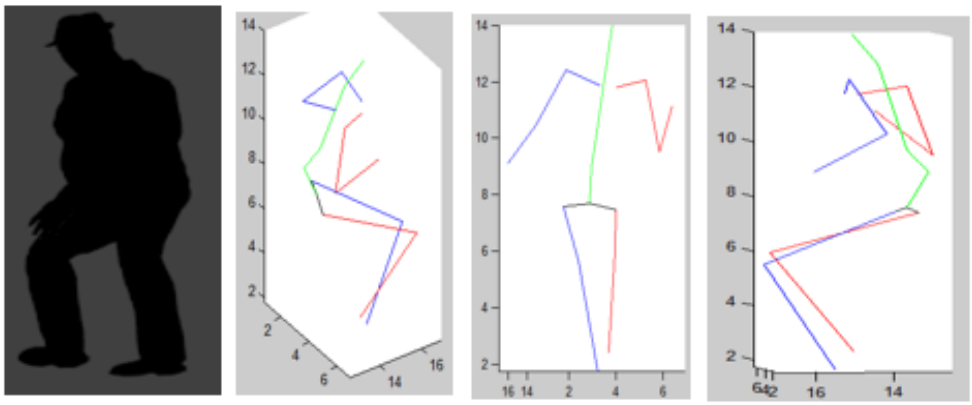


FIGURE A.7 – Résultat de pose 3D avec le descripteur de Hu. A gauche, la silhouette de test, suivi du résultat vu de différents point de vue

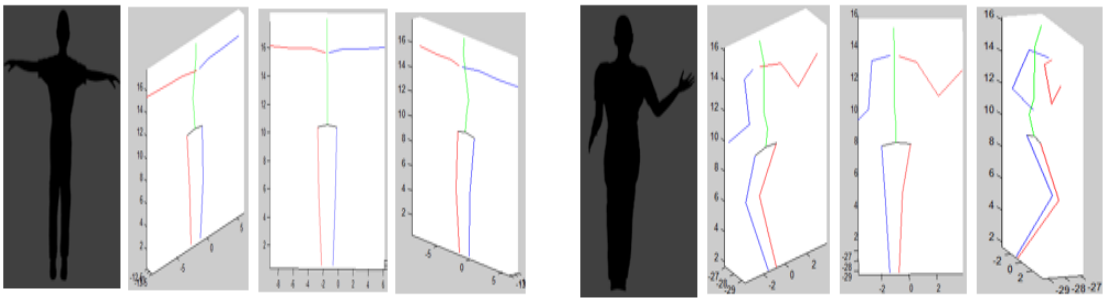


FIGURE A.8 – Résultat de pose 3D avec le descripteur de Krawtchouk : A gauche, la silhouette de test, suivi du résultat vu de différents point de vue

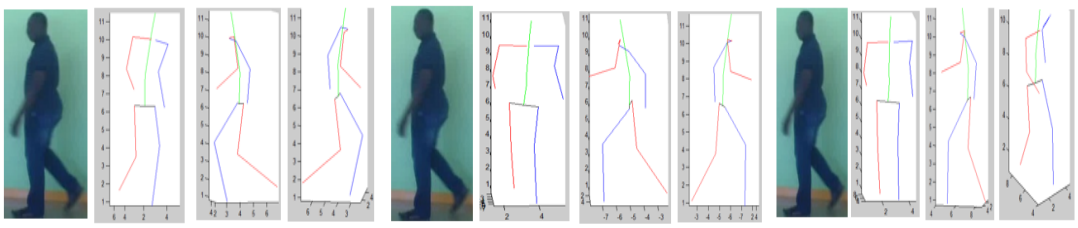
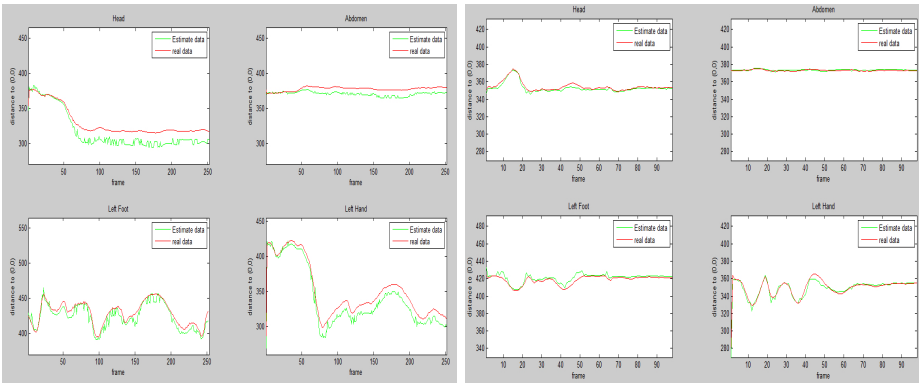


FIGURE A.9 – Résultat de test sur image réel avec respectivement de gauche à droite, les descripteurs Hahn, Krawtchouk et Zernike

un mouvement. Ainsi, la figure A.10 (a), montre le suivi de quatre articulations pour le mouvement "grimper". La courbe rouge est la vérité terrain tandis que la courbe verte représente les données estimées. Les deux courbes présentant des évolutions similaires, on peut dire que la méthode permet d'estimer (pour les données simulées) dans le temps, les articulations des poses, par conséquent, une bonne représentation des formes des silhouettes. La figure (b) représente le suivi des mêmes articulations pour le mouvement de saut.

A.5.3 Evaluation de la robustesse de chaque descripteur au bruit

L'extraction de silhouettes ou plus généralement l'extraction d'objets de premier plan par suppression de l'arrière-plan, n'est pas toujours parfaite et est donc sujet à des bruits. Dans cette sous-section, nous allons évaluer la sensibilité de chaque descripteur face au bruit. Pour ce faire, nous avons utilisé un bruit gaussien appliqué aux silhouettes de la base de test des données



(a) Action "Grimper" (b) Action "Saut"

FIGURE A.10 – Résultats de suivi avec le descripteur de Hahn

simulées afin de les dégrader (voir figA.11). Cette base de test contient 2404 silhouettes n'ayant pas participer dans la partie d'apprentissage. Pour chaque image, soit $x_0 = [0, 0]$ le centre de la silhouette et $x_i = [\rho_i, \theta_i]$ les coordonnées de ses points de contour. L'ajout de bruit $\Delta\sigma$ est fait sur la composante ρ_i avec différentes valeurs d'écart-type : $\Delta\sigma \hookrightarrow \mathcal{N}(0, std)$ avec $std = \{0, 1, 2, 3\}$. Nous rappellons que la base d'apprentissage est composée de 11700 silhouettes.

FIGURE A.11 – Example of noised silhouettes

Les histogrammes de la figure A.12 à la page 117 montre que plus l'écart-type du bruit augmente, plus la performance se dégrade prouvant ainsi que les descripteurs commencent à éprouver des difficultés à trouver les silhouettes correspondantes. Pour le nombre de résultats considéré $N = 1$ avec $std = \{0, 1, 2, 3\}$, les performances sont respectivement $RR = \{35, 44; 35, 44; 24, 95; 18, 96\}$ pour Hu, $RR = \{98, 67; 97; 80, 53; 58, 56\}$ pour Krawtchouk, $RR = \{98, 67; 97, 67; 82, 86; 66, 38\}$ pour Zernike et $RR = \{99, 67; 96; 81, 85; 61, 4\}$ pour Hahn. Ces résultats confirment également que le descripteur avec le moment de Hu n'est pas performant pour cette tâche en comparaison aux trois autres. Le tableau A.1 présente l'erreur moyenne pour chaque descripteur sur l'ensemble de la base de test. En nous basant sur les résultats consignés dans le tableau A.1, nous notons que le descripteur de Hahn dépasse les autres descripteurs si nous considérons plus d'un résultat à la sortie pour la formation du squelette final. En ne considérant qu'un seul résultat, c'est clairement le descripteur de Zernike qui est le plus performant. Bien que nous notons une différence dans les performances des descripteurs, cette différence étant

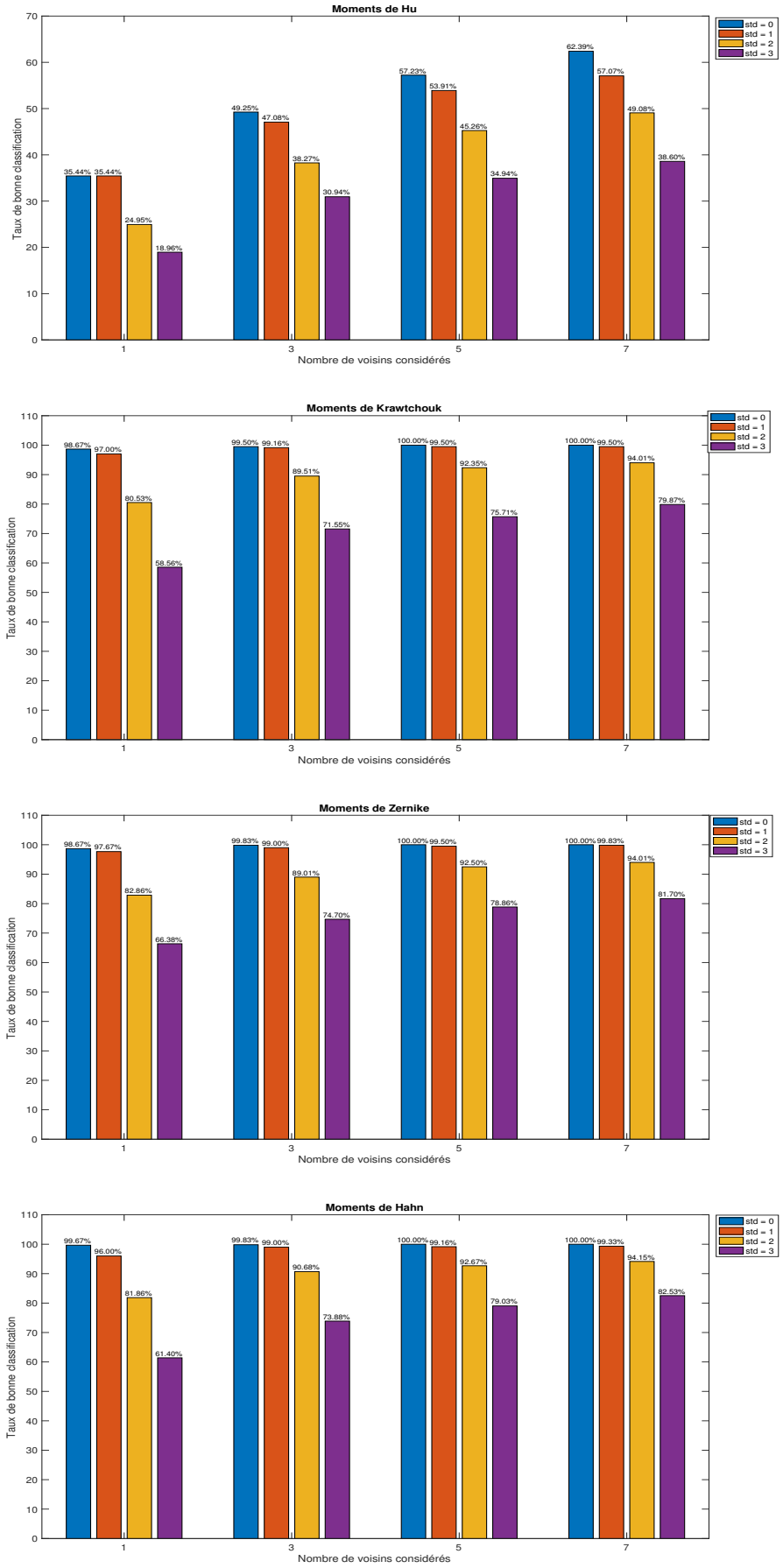


FIGURE A.12 – Histogramme des taux de bonne détection pour les données de test : les couleurs représentent respectivement les résultats pour $std = \{0, 1, 2, 3\}$ pixels.

Tableau A.1 – Résultat récapitulatif de la précision pour chaque descripteur

Descripteurs	**N = 1**	**N = 3**	**N = 5**	**N = 7**
Hu	28,69%	41,38%	47,83%	51,78%
Krawtchouk	83,69%	89,93%	91,89%	93,34%
Zernike	**86,39%**	90,63%	92,71%	93,88%
Hahn	84,73%	**90,84%**	**92,71%**	**94%**

trop faible, une conclusion sur le meilleur moment n'est pas tout à fait possible. On peut en comparant également le temps de calcul, juger du descripteur le plus rapide, mais présentant une performance similaire aux autres. En nous référant au tableau A.2, on note que les descripteurs

Tableau A.2 – Temps d'exécution de chaque descripteur en (s)

Descripteurs	**Hu**	**Kraw**	**Zernike**	**Hahn**
Temps d'exécution	0,047	0,031	0,149	0,036

de Krawtchouk et Hahn présentent les meilleurs temps d'exécution. De la combinaison des ces deux résultats, nous pouvons noter un avantage du descripteur de Hahn sur les trois autres.

Annexe B

L'algorithme EM

L'algorithme EM (Expectation-Maximization en anglais) utilisé pour l'estimation du maximum de vraisemblance des paramètres d'un modèle probabiliste. Il s'agit d'un algorithme itératif composé de deux étapes principales à savoir : une étape d'évaluation de l'espérance et une étape de maximisation de la vraisemblance.

Soit un échantillon d'individus $\mathrm{x} = (x_1, ..., x_n)$ qui suit la loi $f(x_i, \theta)$ paramétrée par θ. L'objectif de l'algorithme est de trouver le paramètre θ qui maximise la log-vraisemblance du modèle (equation B.1).

$$\mathrm{L}(\mathrm{x}; \theta) = \sum_{i=1}^{n} \log f(x_i, \theta) \tag{B.1}$$

En pratique, même au cas où la maximisation de la log-vraisemblance est complexe, il est tout à fait possible de déterminer θ sous réserve de posséder des données bien choisies. Ainsi, des données inconnues dites complètées $\mathrm{z} = (z_1, ..., z_n)$ sont utilisées pour faciliter cette maximisation. La nouvelle log-vraisemblance après intégration des données complètées est donnée par l'équation B.2

$$\mathrm{L}(\mathrm{x}, \mathrm{z}; \theta) = \sum_{i=1}^{n} \Big(\log f(z_i|x_i; \theta) + \log f(x_i; \theta) \Big) \tag{B.2}$$

avec $f(z_i|x_i; \theta)$ la probabilité de z_i sachant x_i et θ.

De l'équation B.2, on a :

$$\mathrm{L}(\mathrm{x}; \theta) = \mathrm{L}(\mathrm{x}, \mathrm{z}; \theta) - \sum_{i=1}^{n} \log f(z_i|x_i; \theta) \tag{B.3}$$

A partir de cette formulation de la log-vraisemblance, l'algorithme EM se base sur l'espérance des données complétées conditionnellement au paramètre courant noté $\theta^{(c)}$.

$$\mathrm{E}\Big[\mathrm{L}(\mathrm{x}; \theta)|\theta^{(c)}\Big] = \mathrm{E}\Big[\mathrm{L}(\mathrm{x}, \mathrm{z}; \theta)|\theta^{(c)}\Big] - \mathrm{E}\Big[\sum_{i=1}^{n} f(z_i|x_i; \theta)|\theta^{(c)}\Big] \tag{B.4}$$

Vu que $\mathrm{L}(\mathrm{x}; \theta)$ ne dépend pas de z, l'équation B.4 peut être réécrite sous la forme :

$$\mathrm{L}(\mathrm{x}; \theta) = \mathrm{Q}\big(\theta; \theta^{(c)}\big) - \mathrm{H}\big(\theta; \theta^{(c)}\big) \tag{B.5}$$

L'algorithme 5 présente les différentes étapes de l'algorithme EM.

Algorithme 5 : Algorithme EM

Entrées : Jeu de paramètre $\theta^{(0)}$ initialisé au hasard
Output : Modèle probabiliste paramétré par $\theta^{(c)}$

c = 0;
tant que *non convergence* **faire**
    Etape E : Evaluation de l'espérance $\mathrm{Q}\big(\theta;\theta^{(c)}\big) = \mathrm{E}\big[\mathrm{L}(\mathrm{x},\mathrm{z};\theta)|\theta^{(c)}\big]$ des données;
    Etape M : Recherche de $\theta^{(c+1)}$ qui maximise la log-vraisemblance
    $\theta^{c+1} = \arg\max_{\theta}\left(\mathrm{Q}\big(\theta;\theta^{(c)}\big)\right)$;
    c = c + 1 ;

L'étape M de l'algorithme fait tendre $\mathrm{L}(\mathrm{x};\theta^{(c+1)})$ vers un maximum local, ce qui représente un inconvénient de l'algorithme. En pratique, pour éviter les maximums locaux, l'algorithme est exécuté un grand nombre de fois avec à chaque fois un jeu initial différent. L'algorithme LDA utilisé dans le chapitre 3 pour chercher le couple (α, β) qui maximise la log-vraissemblance du jeu d'apprentissage. L'étape E de l'algorithme fait une inférence variationnelle avec les paramètres α et β obtenus à l'itération précédente pour calculer la log-vraisemblance tandis que l'étape M cherche à maximiser la borne inférieure de cette log-vraisemblance en α et β.

Publications

Les travaux menés au cours de cette thèse ont été publiés dans des conférences nationales et internationales ainsi que dans des revues internationales. Nous présentons ci-dessous une liste exhaustive de ces différentes publications.

Revues Internationales

— Fabrice Dieudonné ATREVI, Damien VIVET, Florent DUCULTY et Bruno EMILE. A very simple framework for 3D human poses estimation using a single 2D image : comparison of geometric moments descriptors. Pattern Recognition - Volume 71, pages 389-401

Conférences Internationales

— Fabrice Dieudonné ATREVI, Damien VIVET, Florent DUCULTY et Bruno EMILE. "3D Human Poses Estimation from a Single 2D Silhouette". 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Rome, Février 2016.

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Bayesian Generative Model based on Color Histogram Of Oriented Phase and Histogram Of Oriented Optical Flow for Rare Event Detection in Crowded Scenes". 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Avril 2018.

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Rare Events Detection and Localization In Crowded Scenes Based On Flow Signature". 9th IEEE International Conference on Image Processing Theory, Tools and Application (IPTA), Istanbul, Novembre 2019.

Conférences Nationales

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Intégration de la saillance visuelle dans la reconnaissance d'évènements rares". XXVIème Colloque GRETSI, Juan-les-Pins, Septembre 2017.

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Détection d'événements rares par modélisation de la signature du flot optique". XXVIIème Colloque GRETSI, Lille, Août 2019.

Vulgarisations Scientifiques

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Détection et Analyse d'événements rares par vision". Séminaire du laboratoire PRSIME, Orléans, Juin 2017 (Poster).

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Abnormal events analysis by the Latent Dirichlet Allocation (LDA)". International Computer Vision Summer School (ICVSS), Sicile, Juillet 2017 (Poster).

— Fabrice Dieudonné ATREVI, Damien VIVET et Bruno EMILE. "Un modèle bayésien pour l'analyse de mouvement de foule". Séminaire doctorants du LIMOS,Clermond-Ferand, Mai 2018 (Oral).

Bibliographie

[1] Mohiuddin Ahmad and Seong-Whan Lee. Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7) :2237–2252, 2008.

[2] Kevin C Smith, Pedro Quelhas, and Daniel Gatica-Perez. Detecting abandoned luggage items in a public space. Technical report, IDIAP, 2006.

[3] Medha Bhargava, Chia-Chih Chen, Michael S Ryoo, and Jake K Aggarwal. Detection of object abandonment using temporal logic. *Machine Vision and Applications*, 20(5) :271–281, 2009.

[4] Nir Friedman and Stuart Russell. Image segmentation in video sequences : A probabilistic approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 175–181. Morgan Kaufmann Publishers Inc., 1997.

[5] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2458–2465. IEEE, 2009.

[6] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3) :555–560, 2008.

[7] Dieudonné Atrevi, Damien Vivet, and Bruno Emile. Intégration de la saillance visuelle dans la reconnaissance d'événements rares. In *GRETSI*, 2017.

[8] Fabrice Atrevi, Damien Vivet, and Bruno Emile. Bayesian generative model based on color histogram of oriented phase and histogramm of oriented optical flow for rare event detection in crowded scenes. In *43th International Conference on Acoustics, Speech and Signal Processing*, 2018.

[9] Dieudonné Atrevi, Damien Vivet, and Bruno Emile. Détection d'événements rares par modélisation de la signature du flot optique. In *GRETSI*, 2019.

[10] Fabrice Atrevi, Damien Vivet, and Bruno Emile. Rare events detection and localization in crowded scenes based on flow signature. In *9th IEEE International Conference on Image Processing Theory, Tools and Application*, 2019.

[11] Fabrice Atrevi, Damien Vivet, Florent Duculty, and Bruno Emile. 3d human poses estimation from a single 2d silhouette. In *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[12] Dieudonné Fabrice Atrevi, Damien Vivet, Florent Duculty, and Bruno Emile. A very simple framework for 3d human poses estimation using a single 2d image : comparison of geometric moments descriptors. *Pattern Recognition*, 71 :389–401, 2017.

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

[14] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE, 1998.

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110, 2004.

[16] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf : Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

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

[18] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[19] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59, 1996.

[20] Yadong Mu, Shuicheng Yan, Yi Liu, Thomas Huang, and Bingfeng Zhou. Discriminative local binary patterns for human detection in personal album. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[21] Hussin K Ragb and Vijayan K Asari. Color and local phase based descriptor for human detection. In *Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), 2016 IEEE National*, pages 68–73. IEEE, 2016.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1) :142–158, 2016.

[23] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[25] Robert E Schapire. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406, 1999.

[26] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1) :119–139, 1997.

[27] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22, 1977.

[28] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962.

[29] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[30] Zhe Lin, Larry S Davis, David Doermann, and Daniel DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[31] Huchuan Lu, Chunhua Jia, and Ruijuan Zhang. An effective method for detection and segmentation of the body of human in the view of a single stationary camera. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. Citeseer, 2008.

[32] Senjian An, Patrick Peursum, Wanquan Liu, Svetha Venkatesh, and Xiaoming Chen. Exploiting monge structures in optimum subwindow search. 2010.

[33] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2) :154–171, 2013.

[34] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2) :167–181, 2004.

[35] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[36] Iván González-Díaz, Vincent Buso, and Jenny Benois-Pineau. Perceptual modeling in the problem of active object recognition in visual scenes. *Pattern Recognition*, 56 :129–141, 2016.

[37] Shayhan Ameen Chowdhury, Mohammed Nasir Uddin, Mir Md Saki Kowsar, and Kaushik Deb. Occlusion handling and human detection based on histogram of oriented gradients

for automatic video surveillance. In *Innovations in Science, Engineering and Technology (ICISET), International Conference on*, pages 1–4. IEEE, 2016.

[38] Xiao Pu, Xiaoshuang Shi, Zhenhua Guo, and Jie Zhou. Fast human detection using lda via l1-norm. In *Service Sciences (ICSS), 2014 International Conference on*, pages 206–209. IEEE, 2014.

[39] Frédéric Suard, Alain Rakotomamonjy, Abdelaziz Bensrhair, and Alberto Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 206–212. IEEE, 2006.

[40] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2903–2910. IEEE, 2012.

[41] Wei-Ta Chu and Ming-Hung Hsu. Fast object detection using multistage particle window deformable part model. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 98–101. IEEE, 2014.

[42] Sheng Yang, Xian-mei Liao, and UK Borasy. A pedestrian detection method based on the hog-lbp feature and gentle adaboost. *International Journal of Advancements in Computing Technology*, 4(19) :553–560, 2012.

[43] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009.

[44] Yunsheng Jiang and Jinwen Ma. Combination features and models for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 240–248, 2015.

[45] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[46] Noah Sulman, Thomas Sanocki, Dmitry Goldgof, and Rangachar Kasturi. How effective is human video surveillance performance ? In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–3. IEEE, 2008.

[47] Imran Saleemi, Lance Hartung, and Mubarak Shah. Scene understanding by statistical modeling of motion patterns. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2069–2076. IEEE, 2010.

[48] Yang Yang, Jingen Liu, and Mubarak Shah. Video scene understanding using multi-scale analysis. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1669–1676. IEEE, 2009.

[49] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, Dan Xie, Tieniu Tan, and Steve Maybank. A system for learning statistical motion patterns. *IEEE transactions on pattern analysis and machine intelligence*, 28(9) :1450–1464, 2006.

[50] Roger L Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B : Methodological*, 36(6) :507–535, 2002.

[51] Adrien Treuille, Seth Cooper, and Zoran Popović. Continuum crowds. *ACM Transactions on Graphics (TOG)*, 25(3) :1160–1168, 2006.

[52] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009.

[53] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5) :4282, 1995.

[54] Dirk Helbing, Illés Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803) :487, 2000.

[55] Dirk Helbing, Anders Johansson, and Habib Zein Al-Abideen. Dynamics of crowd disasters : An empirical study. *Physical review E*, 75(4) :046109, 2007.

[56] Wenjian Yu and Anders Johansson. Modeling crowd turbulence by many-particle simulations. *Physical review E*, 76(4) :046105, 2007.

[57] Taras I Lakoba, David J Kaup, and Neal M Finkelstein. Modifications of the helbing-molnar-farkas-vicsek social force model for pedestrian evolution. *Simulation*, 81(5) :339–352, 2005.

[58] Hong Liu, Bin Xu, Dianjie Lu, and Guijuan Zhang. A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm. *Applied Soft Computing*, 68 :360–376, 2018.

[59] Yanbin Han and Hong Liu. Modified social force model based on information transmission toward crowd evacuation simulation. *Physica A : Statistical Mechanics and its Applications*, 469 :499–509, 2017.

[60] Nuria Pelechano, Jan M Allbeck, and Norman I Badler. Controlling individual agents in high-density crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 99–108. Eurographics Association, 2007.

[61] Wei Shao and Demetri Terzopoulos. Autonomous pedestrians. *Graphical models*, 69(5-6) :246–274, 2007.

[62] Ramin Mehran, Brian E Moore, and Mubarak Shah. A streakline representation of flow in crowded scenes. In *European conference on computer vision*, pages 439–452. Springer, 2010.

[63] Yassine Benabbas, Nacim Ihaddadene, and Chaabane Djeraba. Motion pattern extraction and event detection for automatic visual surveillance. *Journal on Image and Video Processing*, 2011 :7, 2011.

[64] Tian Wang and Hichem Snoussi. Detection of abnormal visual events via global optical flow orientation histogram. *IEEE Transactions on Information Forensics and Security*, 9(6) :988–998, 2014.

[65] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3449–3456. IEEE, 2011.

[66] Yinghuan Shi, Yang Gao, and Ruili Wang. Real-time abnormal event detection in complicated scenes. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3653–3656. IEEE, 2010.

[67] Dinesh Singh and C Krishna Mohan. Graph formulation of video activities for abnormal activity recognition. *Pattern Recognition*, 65 :265–272, 2017.

[68] Zhijun Fang, Fengchang Fei, Yuming Fang, Changhoon Lee, Naixue Xiong, Lei Shu, and Sheng Chen. Abnormal event detection in crowded scenes based on deep learning. *Multimedia Tools and Applications*, 75(22) :14617–14639, 2016.

[69] Tianlong Bao, Saleem Karmoshi, Chunhui Ding, and Ming Zhu. Abnormal event detection and localization in crowded scenes based on pcanet. *Multimedia Tools and Applications*, 76(22) :23213–23224, 2017.

[70] S Wu, B Moore, and M Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 2054–2060. San Fransisco CA, 2010.

[71] Chen Change Loy, Tao Xiang, and Shaogang Gong. Salient motion detection in crowded scenes. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 1–4. IEEE, 2012.

[72] R Raghavendra, Alessio Del Bue, Marco Cristani, and Vittorio Murino. Abnormal crowd behavior detection by social force optimization. In *International Workshop on Human Behavior Understanding*, pages 134–145. Springer, 2011.

[73] Jing Zhao, Yi Xu, Xiaokang Yang, and Qing Yan. Crowd instability analysis using velocity-field based social force model. In *Visual Communications and Image Processing (VCIP), 2011 IEEE*, pages 1–4. IEEE, 2011.

[74] Hua Yang, Yihua Cao, Shuang Wu, Weiyao Lin, Shibao Zheng, and Zhenghua Yu. Abnormal crowd behavior detection based on local pressure model. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE, 2012.

[75] Wei-Ya Ren, Guo-Hui Li, Jun Chen, and Hao-Zhe Liang. Abnormal crowd behavior detection using behavior entropy model. In *Wavelet Analysis and Pattern Recognition (ICWAPR), 2012 International Conference on*, pages 212–221. IEEE, 2012.

[76] Jaechul Kim and Kristen Grauman. Observe locally, infer globally : a space-time mrf for detecting abnormal activities with incremental updates. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2921–2928. IEEE, 2009.

[77] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. 2009.

[78] Bo Wang, Mao Ye, Xue Li, Fengjuan Zhao, and Jian Ding. Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, 23(3) :501–511, 2012.

[79] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. 2008.

[80] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[81] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1) :18–32, 2014.

[82] Jingxin Xu, Simon Denman, Sridha Sridharan, Clinton Fookes, and Rajib Rana. Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 25–30. ACM, 2011.

[83] Bo Wang, Mao Ye, Xue Li, and Fengjuan Zhao. Abnormal crowd behavior detection using size-adapted spatio-temporal features. *International Journal of Control, Automation and Systems*, 9(5) :905, 2011.

[84] Mehrsan Javan Roshtkhari and Martin D Levine. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer vision and image understanding*, 117(10) :1436–1452, 2013.

[85] Dan Xu, Xinyu Wu, Dezhen Song, Nannan Li, and Yen-Lun Chen. Hierarchical activity discovery within spatio-temporal context for video anomaly detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3597–3601. IEEE, 2013.

[86] Yanjiao Shi, Yunxiang Liu, Qing Zhang, Yugen Yi, and Wenju Li. Saliency-based abnormal event detection in crowded scenes. *Journal of Electronic Imaging*, 25(6) :061608, 2016.

[87] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[89] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2) :36, 2018.

[90] M Sabokrou, M Fathy, and M Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13) :1122–1124, 2016.

[91] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[92] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv :1612.00390*, 2016.

[93] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

[94] Hans P. Morevec. Towards automatic visual obstacle avoidance. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, pages 584–584, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

[95] Stephen M Smith and J Michael Brady. Susan—a new approach to low level image processing. *International journal of computer vision*, 23(1) :45–78, 1997.

[96] Les Kitchen and Azriel Rosenfeld. Gray-level corner detection. *Pattern recognition letters*, 1(2) :95–102, 1982.

[97] Olivier D Faugeras and Marc Berthod. Improving consistency and reducing ambiguity in stochastic labeling : An optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4) :412–424, 1981.

[98] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[99] Hong Zhou, Howard S Friedman, and Rüdiger Von Der Heydt. Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, 20(17) :6594–6611, 2000.

[100] Aude Oliva and Antonio Torralba. Modeling the shape of the scene : A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3) :145–175, 2001.

[101] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis ? *Journal of the ACM (JACM)*, 58(3) :11, 2011.

[102] Xiaodi Hou and Liqing Zhang. Saliency detection : A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[103] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5) :529–541, 1981.

[104] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *computer vision and pattern recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009.

[105] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3) :185–203, 1981.

[106] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[107] Diane J Hu. Latent dirichlet allocation for text, images, and music. *University of California, San Diego. Retrieved April*, 26 :2013, 2009.

[108] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[109] David Arthur and Sergei Vassilvitskii. Worst-case and smoothed analysis of the icp algorithm, with an application to the k-means method. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 153–164. IEEE, 2006.

[110] Edward W Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *biometrics*, 21 :768–769, 1965.

[111] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[112] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

[113] David M Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55(4) :77–84, 2011.

[114] UMN. Unusual crowd activity dataset of university of minnesota, department of computer science and engineering. In *http ://mha.cs.umn.edu/movies/crowd-activity-all.avi*, 2006.

[115] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis : An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018.

[116] Ang Li, Zhenjiang Miao, Yigang Cen, and Qinghua Liang. Abnormal event detection based on sparse reconstruction in crowded scenes. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 1786–1790. IEEE, 2016.

[117] Hajer Fradi, Bertrand Luvison, and Quoc-Cuong Pham. Crowd behavior analysis using local mid-level visual descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 27(3) :589–602, 2017.

[118] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature : Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, 34(1) :194–201, 2012.

[119] Christoph Loeffler, Adriaan Ligtenberg, and George S Moschytz. Practical fast 1-d dct algorithms with 11 multiplications. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 988–991. IEEE, 1989.

[120] Benjamin Heyne, Chi-Chia Sun, Juergen Goetze, and Shanq-Jang Ruan. A computationally efficient high-quality cordic based dct. In *Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE, 2006.

[121] Hung Vu, Dinh Phung, Tu Dinh Nguyen, Anthony Trevors, and Svetha Venkatesh. Energy-based models for video anomaly detection. *arXiv preprint arXiv :1708.05211*, 2017.

[122] Xinfeng Zhang, Su Yang, Xinjian Zhang, Weishan Zhang, and Jiulong Zhang. Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning. *arXiv preprint arXiv :1805.10620*, 2018.

[123] Lei Hu and Fangyu Hu. Anomaly detection in crowded scenes via sa-mhof and sparse combination. In *Computational Intelligence and Design (ISCID), 2017 10th International Symposium on*, volume 1, pages 421–424. IEEE, 2017.

[124] Somaieh Amraee, Abbas Vafaei, Kamal Jamshidi, and Peyman Adibi. Abnormal event detection in crowded scenes using one-class svm. *Signal, Image and Video Processing*, pages 1–9, 2018.

[125] Fabrice Coudert, Jenny Benois-Pineau, and Dominique Barba. Dominant motion estimation and video partitioning with a 1d signal approach. In *Multimedia Storage and Archiving Systems III*, volume 3527, pages 283–295. International Society for Optics and Photonics, 1998.

[126] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *information Theory, IRE Transactions on*, 8(2) :179–187, 1962.

[127] Ramakrishnan Mukundan and KR Ramakrishnan. *Moment functions in image analysis—theory and applications*. World Scientific, 1998.

[128] Michael Reed Teague. Image analysis via the general theory of moments. *JOSA*, 70(8) :920–930, 1980.

[129] Erika Pillu. Analyse et régularisation spatio-temporelle : application à l'écriture manuscrite. Master's thesis, Université Lille1, 2011.

[130] Ramakrishnan Mukundan, SH Ong, and Poh Aun Lee. Image analysis by tchebichef moments. *IEEE Transactions on image Processing*, 10(9) :1357–1364, 2001.

[131] Alireza Khotanzad and Yaw Hua Hong. Invariant image recognition by zernike moments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(5) :489–497, 1990.

[132] Pew-Thian Yap, R. Paramesran, and Seng-Huat Ong. Image analysis by krawtchouk moments. *Image Processing, IEEE Transactions on*, 12(11) :1367–1377, Nov 2003.

[133] Pew-Thian Yap, Raveendran Paramesran, and Seng-Huat Ong. Image analysis using hahn moments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(11) :2057–2062, 2007.

Dieudonné Fabrice ATREVI

Détection et analyse des événements rares par vision, dans un contexte urbain ou péri-urbain

L'objectif principal de cette thèse est le développement de méthodes complètes de détection d'événements rares. Les travaux de cette thèse se résument en deux parties. La première partie est consacrée à l'étude de descripteurs de formes de l'état de l'art. D'une part, la robustesse de certains descripteurs face à différentes conditions de luminosité a été étudiée. D'autre part, les moments géométriques ont été comparés à travers une application d'estimation de pose humaine 3D à partir d'image 2D. De cette étude, nous avons montré qu'à travers une application de recherche de formes, les moments géométriques permettent d'estimer la pose d'une personne à travers une recherche exhaustive dans une base d'apprentissage de poses connues. Cette application peut être utilisée dans un système de reconnaissance d'actions pour une analyse plus fine des événements détectés. Dans la deuxième partie, trois contributions à la détection d'événements rares sont présentées. La première contribution concerne l'élaboration d'une méthode d'analyse globale de scène pour la détection des événements liés aux mouvements de foule. Dans cette approche, la modélisation globale de la scène est faite en nous basant sur des points d'intérêt filtrés à partir de la carte de saillance de la scène. Les caractéristiques exploitées sont l'histogramme des orientations du flot optique et un ensemble de descripteur de formes étudié dans la première partie. L'algorithme LDA (Latent Dirichlet Allocation) est utilisé pour la création des modèles d'événements à partir d'une représentation en document visuel à partir de séquences d'images (clip vidéo). La deuxième contribution consiste en l'élaboration d'une méthode de détection de mouvements saillants ou dominants dans une vidéo. La méthode, totalement non supervisée, s'appuie sur les propriétés de la transformée en cosinus discrète pour analyser les informations du flot optique de la scène afin de mettre en évidence les mouvements saillants. La modélisation locale pour la détection et la localisation des événements est au coeur de la dernière contribution de cette thèse. La méthode se base sur les scores de saillance des mouvements et de l'algorithme SVM dans sa version "one class" pour créer le modèle d'événements. Les méthodes ont été évaluées sur différentes bases publiques et les résultats obtenus sont prometteurs.

Mots clés : Analyse de scènes, détection d'événements rares, allocation latente de dirichlet, mouvements saillants, transformée en cosinus discrète, apprentissage automatique, vision par ordinateur

Rare events detection and analysis by vision, in an urban or peri-urban context

The main objective of this thesis is the development of complete methods for rare events detection. The works can be summarized in two parts. The first part is devoted to the study of shapes descriptors of the state of the art. On the one hand, the robustness of some descriptors to varying light conditions was studied. On the other hand, the ability of geometric moments to describe the human shape was also studied through a 3D human pose estimation application based on 2D images. From this study, we have shown that through a shape retrieval application, geometric moments can be used to estimate a human pose through an exhaustive search in a pose database. This kind of application can be used in human actions recognition system which may be a final step of an event analysis system. In the second part of this report, three main contributions to rare event detection are presented. The first contribution concerns the development of a global scene analysis method for crowd event detection. In this method, global scene modeling is done based on spatiotemporal interest points filtered from the saliency map of the scene. The characteristics used are the histogram of the optical flow orientations and a set of shapes descriptors studied in the first part. The Latent Dirichlet Allocation algorithm is used to create event models by using a visual document representation of image sequences (video clip). The second contribution is the development of a method for salient motions detection in video. This method is totally unsupervised and relies on the properties of the discrete cosine transform to explore the optical flow information of the scene. Local modeling for events detection and localization is at the core of the latest contribution of this thesis. The method is based on the saliency score of movements and one class SVM algorithm to create the events model. The methods have been tested on different public database and the results obtained are promising.

Keywords :Scene analysis, rare event detection, Latent Dirichlet Allocation, salient motion detection, Discrete Cosine Transform, machine learning, computer vision

Laboratoire PRISME, 8 Rue Léonard de Vinci, 45100 Orléans