Vers une nouvelle architecture de videosurveillance basée sur la scalabilité orientée vers l'application

Amal Ben Hamida

THÈSE EN COTUTELLE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR

DE L'UNIVERSITÉ DE BORDEAUX

ET DE L'UNIVERSITÉ DE SFAX
ÉCOLE NATIONALE D'INGENIEURS DE SFAX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES

SPÉCIALITÉ : Informatique

Par Amal BEN HAMIDA

VERS UNE NOUVELLE ARCHITECTURE DE VIDEOSURVEILLANCE BASEE SUR LA SCALABILITE ORIENTEE VERS L'APPLICATION

Sous la direction de Henri NICOLAS
et de Chokri BEN AMAR

Soutenue le: 05 Octobre 2016

Membres du jury:

Mme BEN AZZA, Amel	Professeur	Université de Carthage	Président
M. HAMMAMI, Mohamed	Maître de conférences	Université de Sfax	Rapporteur
Mme. MORIN, Luce	Professeur	INSA de Rennes	Rapporteur
M. ZAGROUBA, Ezzeddine	Professeur	Université de Tunis El Manar	Examinateur
M. BEN AMAR, Chokri	Professeur	Université de Sfax	Directeur
M. NICOLAS, Henri	Professeur	Université de Bordeaux	Directeur

Titre : Vers une nouvelle architecture de vidéosurveillance basée sur la scalabilité orientée vers l'application

Résumé : Le travail présenté dans ce mémoire a pour objectif le développement d'une nouvelle architecture pour les systèmes de vidéosurveillance. Tout d'abord, une étude bibliographique nous a conduit à classer les systèmes existants selon le niveau de leurs applications qui dépend directement des fonctions analytiques exécutées. Nous avons également constaté que les systèmes habituels traitent toutes les données enregistrées alors que réellement une faible partie des scènes sont utiles pour l'analyse. Ainsi, nous avons étendu l'architecture ordinaire des systèmes de surveillance par une phase de pré-analyse qui extrait et simplifie les régions d'intérêt en conservant les caractéristiques importantes. Deux méthodes différentes pour la pré-analyse dans le contexte de la vidéosurveillance ont été proposées : une méthode de filtrage spatio-temporel et une technique de modélisation des objets en mouvement. Nous avons contribué, aussi, par l'introduction du concept de la scalabilité orientée vers l'application à travers une architecture multi-niveaux applicatifs pour les systèmes de surveillance. Les différents niveaux d'applications des systèmes de vidéosurveillance peuvent être atteints incrémentalement pour répondre aux besoins progressifs de l'utilisateur final. Un exemple de système de vidéosurveillance respectant cette architecture et utilisant nos méthodes de pré-analyse est proposé

Mots clés : vidéosurveillance, pré-analyse vidéo, scalabilité orientée vers l'application

---

Title : Towards a new video surveillance architecture based on the application-oriented scalability

Abstract : The work presented in this thesis aims to develop a new architecture for video surveillance systems. Firstly, a literature review has led to classify the existing systems based on their applications level which dependents directly on the performed analytical functions. We, also, noticed that the usual systems treat all captured data while, actually, a small part of the scenes are useful for analysis. Hence, we extended the common architecture of surveillance systems with a pre-analysis phase that extracts and simplifies the regions of interest with keeping the important characteristics. Two different methods for pre-analysis were proposed : a spatio-temporal filtering and a modeling technique for moving objects. We contributed, too, by introducing the concept of application-oriented scalability through a multi-level application architecture for surveillance systems. The different applications levels can be reached incrementally to meet the progressive needs of the end-user. An example of video surveillance system respecting this architecture and using the pre-analysis methods was proposed.

Keywords : video surveillance, video pre-analysis, application-oriented scalability

---

Laboratoire Bordelais de Recherche en Informatique (LABRI)-UMR 5800
Université de Bordeaux
351 Cours de la Libération
33405 Talence Cedex FRANCE

Remerciements

Cette thèse de Doctorat a été réalisée dans le cadre d'une convention de cotutelle entre l'Ecole Nationale d'Ingénieurs de Sfax, Université de Sfax et l'Université de Bordeaux. Les recherches qui font l'objet de ce mémoire ont été réalisées sur deux sites: en France, au sein du Laboratoire Bordelais de Recherche en Informatique (LaBRI) et en Tunisie, dans le Laboratoire Research Groups in Intelligent Machines (ReGIM-Lab).

Par ces quelques lignes, je tiens à remercier toutes les personnes qui ont participé de près ou de loin au bon déroulement de cette thèse, en espérant n'avoir oublié personne…

Je tiens à remercier vivement Mr. Chokri BEN AMAR, Professeur à l'Ecole Nationale d'Ingénieurs de Sfax (ENIS), qui a co-dirigé cette thèse, d'avoir bien assuré la direction et l'encadrement de mes travaux de thèse. Merci pour votre gentillesse, votre patience et vos précieux conseils. Je garde toujours beaucoup de plaisir à discuter avec vous et à bénéficier de vos conseils. Je manque d'expressions de remerciements dignes de tout ce que vous m'avez donné durant ma thèse.

Je tiens à remercier spécialement mon directeur de thèse Mr. Henri NICOLAS, Professeur à l'Université de Bordeaux en France, d'avoir cru en mes capacités, pour le temps et la patience que vous m'avez accordés tout au long de ces années en me fournissant d'excellentes conditions de travail. Votre disponibilité illimitée, vos conseils constructifs et votre suivi minutieux de toutes les particularités de mon travail méritaient d'être ardemment remerciés. J'ai beaucoup appris à vos cotés et je ne parviens jamais à vos adresser les expressions de reconnaissance que vous méritez.

Mes remerciements particuliers s'adressent aussi à Mr. Mohamed KOUBAA, Maitre assistant à l'École Nationale d'Electronique et des Télécommunications (ENET'Com), qui à participer également à l'encadrement de cette thèse, d'avoir suivi de façon régulière le travail pour son aide permanent et ses conseils valeureux. Mr. KOUBAA restera un grand frère avec qui j'apprécie beaucoup travailler tant sur le plan scientifique que sur le plan humain.

Je souhaite exprimer toute ma gratitude envers les membres du jury qui ont accepté de consacrer à cette thèse une partie de leurs temps extrêmement précieux.

Je commence par remercier Mme. Amel BEN AZZA, Professeur à l'École Supérieure des Communications de Tunis (SUP'COM), pour avoir accepté de juger ce travail et d'en présider le jury de soutenance.

Je remercie également Mr. Ezzeddine ZAGROUBA, Professeur à l'Université Virtuelle de Tunis (UVT), qui m'a honoré par sa présence en qualité d'examinateur de ma thèse, pour l'intérêt qu'il a porté à mon travail.

Je remercie en particulier les rapporteurs de ce travail Mme. Luce MORIN, Professeur à l'Institut National des Sciences Appliquées de Rennes (INSA Rennes), et Mr. Mohamed HAMMAMI, Maitre de conférences à la Faculté des Sciences de Sfax (FSS), pour leurs lectures attentives de mon rapport de thèse et leurs remarques pertinentes.

Il convient aussi de remercier profondément tous les membres du laboratoire ReGIM et surtout le Professeur Adel Mohamed ALIMI, directeur du laboratoire ReGIM-Lab, pour ses qualités pédagogiques et scientifiques, sa franchise et ses conseils valeureux.

Je ne dois pas aussi oublier mes amies du LaBRI, Nesrine, Sameh et Omsaad avec qui j'ai passé des agréables moments pendant mes séjours en France.

Bien évidemment, je remercie mes parents et tous les membres de ma famille, en particulier mon mari et ma petite fille ZAINAB pour leur soutien permanent et surtout de m'avoir supporter pendant toutes ces longues et dures années.

Table des matières

Chapitre 2 : Une nouvelle approche de pré-analyse vidéo pour les systèmes de vidéosurveillance

Table des figures

Introduction générale

Contexte général

Les systèmes de la vidéosurveillance jouent un rôle de plus en plus important dans la surveillance à distance des personnes, des biens et des sites publics et privés. Leurs premières apparitions étaient dans les années 1950. Néanmoins, la surveillance a vraiment été développée à partir des années 1970 au moyen des systèmes de télévision en circuit fermé (CCTV), principalement au Royaume-Uni. L'implantation de la vidéosurveillance s'est intensifiée au cours des années 1990. Depuis les attaques de 2001 aux États-Unis et 2005 à Londres, le nombre de systèmes de surveillance installés s'est élevé. Ainsi, un nombre considérable de caméras a été largement déployé dans les espaces publics, y compris les infrastructures de transport (les aéroports, les stations de métro,...), les parkings, les banques, les centres commerciaux, les routes et les sites industriels comme outil de réduction de la criminalité et de la gestion des risques [Gouaillier 2009].

De nos jours, la vidéosurveillance constitue l'une des solutions de sécurité les plus anciennes et répandues. L'apparition des caméras IP a amorcé le passage de la technologie analogique CCTV vers la vidéo sur réseaux IP. Ceci a facilité l'installation des réseaux de vidéosurveillance comptant un grand nombre de caméras, par exemple dans un aéroport, des centaines de caméras de surveillance peuvent être déployées. Ces larges infrastructures de vidéosurveillance mènent à une quantité colossale de flux vidéo à transmettre, visionner et archiver. En même temps, la majorité de ces systèmes comptent beaucoup sur les opérateurs humains pour surveiller les scènes et déceler les comportements ou les événements suspects. Malheureusement, de nombreux incidents ne sont pas détectés en raison de certaines limitations fortement liées aux capacités de la surveillance humaine : (1) plusieurs écrans à regarder à la fois par un même opérateur (en pratique, chaque opérateur humain ne peut pas contrôler plus que 4 écrans à la fois [Wallace 1988]); (2) l'ennui, la fatigue et la monotonie en raison des heures continues de surveillance (une pause de 5-10 minutes est recommandée chaque heure pour des raisons de santé et de sécurité [Wallace 1988]); (3) des informations ambiguës et floues sur ce qu'on cherche dans l'écran (les incidents ne peuvent pas toujours être prédits, ils peuvent se produire de façon inattendue). En même temps, les comportements anormaux déclenchent souvent des soupçons, mais ils ne conduisent pas toujours à des

incidents); (4) le choix de la caméra à regarder est pris par l'opérateur, ce qui rend le système vulnérable aux abus (des études sociologiques affirment que les agents de vidéosurveillance décident fréquemment quelle caméra surveiller en se basant sur l'apparence plutôt que sur le comportement des personnes à l'écran [Smith 2004]); (5) d'autres tâches peuvent être à gérer par l'opérateur humain en plus de la surveillance (l'émission des badges et des clés, le contrôle des communications radio,...); (6) l'honnêteté et le sérieux des opérateurs sont parfois mis en doute.

Dans les larges systèmes de surveillance, des centaines de caméras sont montées pour assurer un contrôle absolu. Basé sur l'étude de [Dee 2008] au Royaume-Uni, le rapport écrans/caméras est compris entre 1/4 et 1/30 et le rapport agents/écrans est d'environ 1/16. Ainsi, bien que théoriquement toutes les caméras soient surveillées, seulement un petit nombre d'écrans sont surveillés en temps réel. Même ces derniers ne peuvent pas être surveillés convenablement à cause des six limitations déjà citées. Le reste est enregistré et regardé par la suite, si nécessaire, il s'agit alors d'une utilisation a posteriori. Pour surmonter ces problèmes, une vague de migration vers les systèmes de surveillance intelligents est en train d'émerger récemment en exploitant les techniques avancées en analyse vidéo et en intelligence artificielle. Le but de l'utilisation des techniques de vision par ordinateur est d'imiter la perception visuelle et l'analyse des humains. Quoique les systèmes de surveillance intelligents surpassent les capacités humaines dans de nombreux cas et les performances de l'homme sont loin d'être optimales, les systèmes de surveillance restent toujours sous la supervision de l'agent humain.

Dans les systèmes de surveillance courants, toute la scène est enregistrée et analysée. Le traitement de toutes ces données alourdit les capacités du système. Prenons l'exemple d'un système de vidéosurveillance industrielle qui compte 14 caméras. Si chacune de ces caméras fonctionne conformément aux standards de la résolution, des paramètres de qualité et des débits d'images, pour une simple surveillance de 24 heures, environ 18,4 milliards de giga-octets de données sont à traiter. Néanmoins, réellement, seule une faible proportion des informations enregistrées est vraiment importante. Elle représente les zones de la scène où les événements d'intérêt se déroulent. Pour éviter cet énorme et inutile traitement des flux de données, nous proposons de détecter, grâce à une phase de pré-analyse au niveau de la caméra (capteur), les zones susceptibles de contenir des informations importantes, puis de les filtrer pour éliminer l'information a priori inutile. Finalement seules ces zones simplifiées sont compressées puis transmises pour être analysées plus finement au niveau de la réception. L'étape de l'analyse finale est très dépendante de l'application finale du système. En fait, les

systèmes de vidéosurveillance ont diverses applications qui peuvent être classées en quatre catégories partant des applications de bas niveau jusqu'à des applications de haut niveau. La catégorie des applications de premier niveau regroupe les systèmes de surveillance destinés à détecter et compter des objets en mouvement dans la scène. La classe des applications de second niveau constitue les systèmes capables de détecter et de suivre les objets en mouvement. Les systèmes appartenant à la troisième catégorie peuvent classer et identifier les objets suivis. La classe des applications de quatrième niveau contient les systèmes avec des applications plus avancées visant à analyser et à comprendre les comportements des objets en mouvement dans la scène. Le niveau applicatif du système est très dépendant de l'application désirée par l'utilisateur final, ce qui permet d'identifier les tâches à accomplir au cours de l'étape de l'analyse finale. Jusqu'à présent, les systèmes de vidéosurveillance déjà existants sont à un seul niveau d'application : ils sont conçus pour répondre à une seule application spécifique bien déterminée. Notre idée est de proposer un nouveau type d'architecture qui sera à plusieurs niveaux d'application, contrairement à l'architecture habituelle (à un seul niveau). En vertu de cette architecture, le niveau de l'application du système peut être amélioré à la demande de l'utilisateur final. Le traitement des données se fait de manière scalable. Nous nous intéressons aussi au cours de cette thèse à la scalabilité, qui n'a été utilisée jusqu'à maintenant que pour le codage des flux vidéo. En effet, nous proposons un concept de scalabilité orientée vers l'application pour des objectifs de surveillance qui assure, suivant l'application du système, l'extraction et le traitement des informations pertinentes à partir des données enregistrées. Au cas où l'utilisateur final demanderait une application de niveau plus élevé, les caractéristiques requises pour répondre à ses besoins sont extraites et traitées. Le schéma de notre architecture proposée est : enregistrer, pré-analyser, compresser, transmettre, décompresser et analyser la scène. Elle fonctionne de manière scalable pour s'aligner avec les besoins progressifs de l'utilisateur final. Elle est appliquée dans le contexte de la vidéosurveillance routière et des lieux publics.

Contributions

Au cours de cette thèse, trois importantes constations ont mené à développer des contributions dans le domaine de la vidéosurveillance routière et des lieux publics. Premièrement, nous avons constaté que les systèmes de vidéosurveillance servent plusieurs domaines applicatifs. Ces diverses applications sont réalisées en se basant sur une ou plusieurs tâches analytiques et ont des objectifs différents en fonction des besoins de

l'utilisateur final. Par conséquent, nous avons proposé une nouvelle classification pour les systèmes de vidéosurveillance suivant un critère relié au niveau de l'application. En deuxième lieu, nous avons remarqué que seulement une faible partie des données enregistrées est vraiment utile. Ainsi, nous avons proposé d'ajouter une étape de pré-analyse à l'architecture habituelle des systèmes de vidéosurveillance afin d'envoyer uniquement les informations pertinentes pour l'analyse finale. Nous avons élaboré deux méthodes de pré-analyse qui détectent les régions d'intérêt dans la séquence et qui les simplifient sans détruire les informations d'intérêt qu'elles emportent. Troisièmement, le concept de la scalabilité dans le contexte de la vidéosurveillance n'a été abordé que dans la phase de codage de flux vidéo. Dans ce travail, nous avons appliqué le concept de la scalabilité dans l'aspect applicatif des systèmes de surveillance. La scalabilité est intégrée pour éviter la transmission et l'analyse des flux de données inutiles en tenant compte du niveau applicatif du système. Ainsi, une architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance est proposée. L'objectif général de cette thèse est de fournir une nouvelle architecture qui évite d'accabler les systèmes de surveillance en traitant toutes les données enregistrées. Selon cette architecture, les informations pertinentes sont progressivement extraites, envoyées et traitées suivant les besoins.

Le travail présenté dans ce mémoire apporte quatre contributions majeures dans le domaine de la vidéosurveillance :

1. *Classification des systèmes de vidéosurveillance suivant le niveau de l'application.* Les systèmes de vidéosurveillance servent plusieurs domaines. Cette multitude d'applications peut être divisée en quatre classes partant des applications de bas niveau jusqu'à des applications de haut niveau suivant les tâches exécutées.

2. *Filtrage spatio-temporel des objets en mouvement.* Une première méthode de pré-analyse des données qui permet de simplifier la séquence vidéo dans les deux dimensions : spatiale et temporelle. Elle permet de ne garder que les détails nécessaires pour une analyse ultérieure.

3. *Modélisation géométrique des objets en mouvement.* Une deuxième méthode de pré-analyse des données qui simplifie la représentation de l'objet en une forme parallélépipédique, dans le plan de l'image, avec le maintien des caractéristiques requises pour le suivi d'objets.

4. *Architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance.* Cette approche gère la scalabilité du point de vue applicatif. Elle affecte le niveau de l'application en permettant à tout système de vidéosurveillance d'être adaptable et progressif relativement à sa réponse à l'utilisateur final. L'architecture proposée a une structure hiérarchique à différents niveaux. Un exemple de système de vidéosurveillance qui respecte cette architecture est proposé. Pour chaque niveau, les caractéristiques demandées sont extraites, envoyées et analysées en fonction de l'application souhaitée. Les deux méthodes de pré-analyse développées sont utilisées dans cet exemple.

Organisation du mémoire

Le présent rapport est structuré en trois chapitres et s'organise de la façon suivante :

Le chapitre 1 dresse un état de l'art sur les différentes notions de la vidéosurveillance adressées dans cette thèse. Sa première partie présente certaines généralités sur la vidéosurveillance explicitant l'importance de cette solution de sécurité à l'échelle mondiale et citant quelques systèmes répandus de vidéosurveillance. Elle décrit aussi l'architecture habituelle et les classifications existantes des systèmes de vidéosurveillance. Nous y proposons une nouvelle classification possible des systèmes de vidéosurveillance qui est : la classification suivant le niveau d'application. La deuxième partie présente les méthodes couramment utilisées dans l'analytique vidéo et cite, de manière non exhaustive, des techniques de détection d'objets, de suivi d'objets, de classification d'objets et d'analyse de comportements. Quant aux troisième et quatrième parties respectivement, elles présentent l'apport de la modélisation d'objets en mouvement et la scalabilité dans le contexte de la vidéosurveillance. A l'issue de cette étude des travaux de l'état de l'art, la sixième section dégage les limitations qui motivent le travail de thèse.

Le chapitre 2 s'intéresse à l'approche de pré-analyse vidéo pour les systèmes de vidéosurveillance. La chaine complète et les étapes de traitement sont présentées en premier lieu. Ensuite, les descriptions de la méthode de filtrage spatio-temporel et de la technique de modélisation géométrique d'objets en mouvement sont respectivement détaillées dans les quatrième et cinquième parties. A l'issue des présentations des contributions, les évaluations expérimentales des performances, dans le cadre de la surveillance routière et des lieux publics, sont exposées pour la validation.

Le chapitre 3 décrit, en premier lieu, l'architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance. Ensuite, un exemple de système de surveillance respectant cette architecture est présenté. Il est utilisé dans l'étude expérimentale dans le but de valider l'approche proposée ainsi que de voir ses limites.

En conclusion, nous faisons une synthèse sur les apports et les limitations des contributions présentées dans ce mémoire, et nous clôturons par la proposition de certaines perspectives de recherche.

Chapitre 1

Etat de l'art sur la vidéosurveillance

1. Introduction

La vidéosurveillance est un segment de l'industrie de la sécurité physique qui consiste à surveiller à distance des lieux publics ou privés, à l'aide de caméras. Les caméras de surveillance sont peu coûteuses et partout de nos jours, mais la main-d'œuvre nécessaire pour surveiller et analyser est chère. Par conséquent, les flux vidéos de ces caméras sont généralement surveillés avec modération ou ignorés. Ils sont souvent utilisés comme de simples archives ou pour renvoyer une alerte, une fois un incident a eu lieu. Aujourd'hui, les caméras de surveillance sont devenues un outil beaucoup plus utile. Au lieu de passivement enregistrer des images, elles sont utilisées pour détecter des événements nécessitant une attention en même temps qu'ils se produisent, et prendre des mesures en temps réel. La vidéosurveillance pour l'homme est l'un des sujets de recherche les plus actifs dans la vision par ordinateur. Elle dispose d'une variété d'applications de sécurité prometteuse.

Ce chapitre expose un état de l'art sur la vidéosurveillance. Dans une première section, il montre sa situation mondiale actuelle, ses projets les plus répandus et l'architecture de ses systèmes. Il dégage aussi les types des systèmes de vidéosurveillance, ainsi que leurs applications et décrit une nouvelle classification basée sur le niveau applicatif du système. La Section 3 fournit un état sur l'avancement technologique en analytique vidéo tout en décrivant les techniques d'analytique vidéo dans les systèmes de la vidéosurveillance. La Section 4 présente les apports de la modélisation des objets pour la vidéosurveillance en citant des travaux connexes. La Section 5 décrit le rôle actuel de la scalabilité dans le contexte de la vidéosurveillance. Enfin, la Section 6 discute les limitations des travaux de l'état de l'art pour annoncer les motivations du travail de thèse.

2. Généralités sur la vidéosurveillance

2.1. *Situation mondiale de la vidéosurveillance*

De nos jours, nous sommes les témoins de la prolifération de la vidéosurveillance dans tous les pays du monde. Les systèmes de vidéosurveillance commercialisés servent plusieurs applications : la protection des sites sensibles (les édifices gouvernementaux, les centrales nucléaires, les barrages fluviaux), et des lieux publics (les musées, les aéroports, les gares, les banques, les centres commerciaux), la sécurité des domiciles (détection de vol, détection d'incendie), la surveillance des personnes âgées (analyse d'activités, détection de chute), la

sécurité routière (estimation de flux, contrôle du trafic aérien, détection d'accidents), la détection d'événements anormaux (détection de tricherie dans les écoles, détection de criminalité dans les villes), la sécurité industrielle (surveillance des travailleurs dans les manufactures, contrôle d'accès), etc [Gong 2011].

Selon l'étude de IHS Technology, 245 millions de caméras de vidéosurveillance installées de façon professionnelle sont actives et opérationnelles à l'échelle mondiale en 2014 [IHS 2014]. Les 5 premières villes du monde contenant plus de caméras de surveillance sont respectivement : Chongqing (Chine), Beijing (Chine), London (Royaume-Uni), Chicago (Etats-Unis), Houston (Etats-Unis), New York City (Etats-Unis). La sécurité nationale, étant devenue une affaire très délicate, est le plus important moteur du marché de la vidéosurveillance. La Figure 1 illustre les dépenses, en milliards de Dollars, en sécurité nationale de quelques pays qui ne cessent de s'amplifier [Gouaillier 2009]. Parallèlement, le marché de la surveillance est en pleine croissance, avec une progression de 8 à 12 % par an, comme le montre la Figure 2. Le taux de croissance du marché mondial en 2014 est de 12% pour atteindre 15,9 milliards de Dollars contre 14,1 milliards en 2013 [Protection 2014]. Le marché mondial de la vidéosurveillance IP devrait atteindre près de 43 milliards d'euros en 2019 [Surveillance 2013]. Le Royaume-Uni, connu comme "l'état de la surveillance", est le leader européen avec le plus grand nombre de caméras [Norris 2004]. En 2013, le BSIA (British Security Industry Association) a estimé qu'il y a 5,9 millions de caméras dans le pays qui est d'environ un pour chaque 11 personnes [BSIA 2013]. La France a également adopté massivement les solutions technologiques de surveillance. Selon une enquête menée par INSEE (Institut National de la Statistique et des Études Économiques), 100% des villes françaises de droite sont équipées de systèmes de vidéosurveillance avec en moyenne 1858 habitants par caméra et 60% des villes de gauche sont équipées avec une caméra pour chaque 4961 habitants [Palmares 2011]. Des augmentations de l'utilisation de la vidéosurveillance sont également décrites au Moyen-Orient, en Afrique du Sud, en Australie, en Inde, en Russie et en Europe orientale. La Tunisie fait partie des pays qui ont un besoin urgent des solutions de la vidéosurveillance. L'INPDP (Instance Nationale de Protection des Données à Caractère Personnel) ne dispose pas de chiffres précis sur le nombre de caméras de vidéosurveillance installées sur l’ensemble du territoire tunisien, bien que les chiffres augmentent. Elle estime cependant que seulement 10% d’entre elles ont obtenu la licence préalable à leur utilisation.

2.2. *Systèmes répandus de la vidéosurveillance*

La nécessité de la vidéosurveillance a mené au lancement de différents grands projets de recherche. Plusieurs d'entre eux ont prouvé leurs efficacités et sont devenus des solutions mondiales très répandues, citons : le système VSAM (Video Surveillance and Monitoring) qui permet d'analyser automatiquement les activités des objets dans les champs de batailles ou les scènes civiles ordinaires [Collins 2000]; le système Pfinder réalisant un suivi précis d'une personne en mouvement dans des scènes complexes; la solution S3 (Smart Surveillance Solution) de IBM qui permet la détection, le suivi et la classification d'objets selon la couleur du visage [Corporation 2009]; le système de surveillance en temps réel W4 qui utilise une combinaison de l'analyse de la forme et de suivi, et construit des modèles d'apparitions des personnes afin de détecter, suivre les groupes de personnes en occlusion et surveiller leurs comportements [Haritaoglu 2000]; le système HID (Human Identification at a Distance) qui classifie et identifie les êtres humains à grandes distances; le projet européen ADVISOR (AiDe à la VIdéoSurveillance) qui est un projet de base sur la surveillance dans les stations de métro; le système de surveillance routière VIEWS qui joue un rôle très important dans le contrôle du trafic [Javed 2003]; Smart Catch utilisé dans l'aéroport international de San Francisco peut détecter les anomalies ou les comportements suspects; IVA software (Intelligent Video Analysis) qui assure la surveillance de l'aéroport international d'Athènes [Bosh 2001].

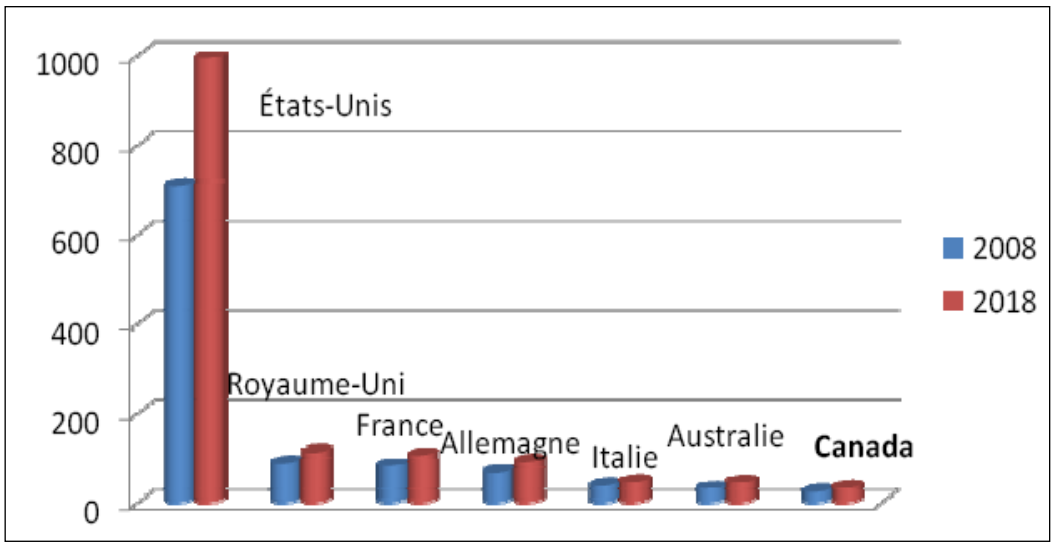

Figure 1. Croissance des dépenses en sécurité nationale de quelques pays entre 2008 et 2018, en milliards de Dollars [Gouaillier 2009]

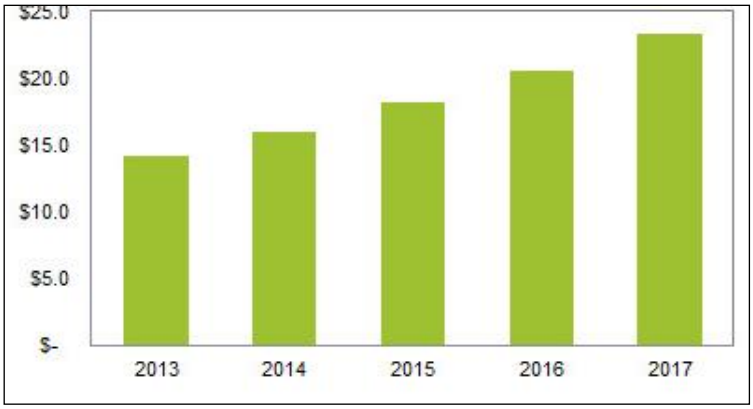

Figure 2. Revenus du marché mondial du CCTV en milliards de Dollars [IHS 2014]

2.3. *Architecture des systèmes de vidéosurveillance*

Dans cette section, on présente de façon sommaire les différentes composantes matérielles et logicielles des systèmes de vidéosurveillance. Comme l'illustre la Figure 3, les systèmes de vidéosurveillance courants sont généralement composés des étapes suivantes : l'acquisition, la compression, la transmission, la décompression et le traitement.

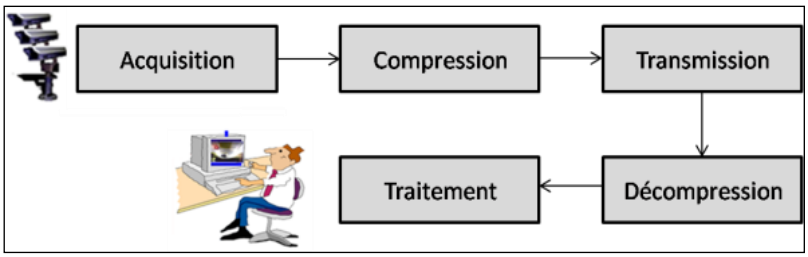

Figure 3. Architecture générale des systèmes de vidéosurveillance

- *Acquisition.* La scène à contrôler est enregistrée par une caméra de surveillance. Il existe une variété de modèles de caméras répondant aux différents besoins de surveillance. Elles sont analogiques ou numériques et peuvent être motorisées ou non.

- *Compression.* La séquence vidéo numérisée représente une grande quantité de données à transmettre et à traiter. Ce qui nécessite une bande passante assez large et un espace de stockage important. Puisque ceci n'est pas toujours disponible, la vidéo doit être compressée afin de réduire la quantité de données en supprimant les redondances entre les images ainsi que les détails imperceptibles à l'œil humain.

- *Transmission.* La séquence vidéo captée par les caméras de surveillance doit être transmise aux stations de traitement et d'enregistrement. Plusieurs moyens de transmission sont fournis.
- *Traitement.* A leurs arrivées à la station finale, les flux vidéos peuvent subir de différents traitements, suivant l'objectif de l'application du système de surveillance, tels que l'enregistrement, le visionnement, l'analyse et la recherche dans les séquences enregistrées. Certains systèmes archivent simplement les séquences vidéo pour une durée limitée. Les enregistrements ne sont visionnés qu'en cas de besoin. D'autres assurent une supervision directe des centaines de caméras par des opérateurs humains. Les systèmes de vidéosurveillance intelligents analysent automatiquement en temps réel les scènes transmises et alertent l'opérateur en cas de soupçon.

2.4. *Classifications des systèmes de vidéosurveillance*

Une grande variété de systèmes de surveillance est proposée jusqu'à aujourd'hui. Ces systèmes peuvent être classés en fonction de différents critères.

2.4.1. Classification suivant le niveau d'automatisation

Après l'exploitation des techniques avancées d'analyse vidéo et d'intelligence artificielle, les systèmes de vidéosurveillance peuvent être classés en trois types : manuels, semi-autonomes et entièrement autonomes [Gouaillier 2009].

- *Les systèmes de surveillance manuels.* Ils impliquent avoir un opérateur humain surveillant les écrans directement ou utilisant les enregistrements de suite. Ces systèmes sont encore largement utilisés.

- *Les systèmes de surveillance semi-autonomes.* Ils combinent le traitement vidéo et l'intervention humaine. Prenons l'exemple d'un système où seulement les mouvements imprévus sont enregistrés et envoyés pour une analyse par un expert humain.

- *Les systèmes de vidéosurveillance autonomes.* Ils sont aussi appelés les systèmes de la vidéosurveillance intelligente. Ils peuvent assurer un suivi fiable en temps réel en analysant intelligemment les données vidéo sans l'intervention de l'homme. Les systèmes intelligents doivent remplir trois caractéristiques importantes : (1) l'utilisation sans le contrôle humain; (2) la prédiction des événements, des

comportements, des mouvements; (3) le suivi, le contrôle et l'alerte en cas d'activités imprévues.

2.4.2. Classification suivant l'architecture réseau

Les systèmes de vidéosurveillance peuvent être déployés selon deux grands types d'architecture, soit centralisée ou distribuée [Gouaillier 2009].

- *Architecture centralisée*. Dans une architecture centralisée, tous les traitements sont effectués dans la même station de contrôle. L'encodage, l'enregistrement, le visionnement et l'analyse des flux vidéo nécessitent une grande puissance de calcul. De plus, la transmission de tous les flux vidéo en un point centralisé consomme beaucoup en bande passante.

- *Architecture distribuée.* Dans une architecture distribuée les traitements sont répartis dans les différents nœuds du système de vidéosurveillance. Ainsi, les calculs nécessaires à l'analyse peuvent être faits sur des caméras intelligentes dotées de processeurs, ou dans les encodeurs. Cette architecture réduit la bande passante nécessaire et facilite l'extension du réseau de caméras puisque l'ajout de caméras n'affecte pas la puissance de calcul de la station finale.

2.4.3. Classification suivant le domaine d'application

Les applications de vidéosurveillance peuvent être divisées en cinq catégories en fonction de leurs objectifs [Chamasemani 2013].

- *Protection et confidentialité.* La vidéosurveillance est massivement déployée pour la protection des personnes et des lieux. Elle est largement utilisée par les gouvernements pour la sécurité intérieure [Ko 2008], et la sécurité des sites publics (musée, aéroport, gare, banque, ...) [Dimitropoulos 2009, Sehchan 2007]. Elle sert également à la surveillance des domiciles [Bai 2011, Loomans 2011], la surveillance des personnes âgées [Chua 2015, Wang 2015], ...

- *Analyse de l'objet.* Certains systèmes de surveillance vidéo sont utiles pour découvrir les trajectoires des personnes par le suivi [Ezzahout 2013, Peng 2014]. D'autres aident à surveiller des environnements complexes comme la supervision des activités des travailleurs dans les manufactures [Fuhai 2012], l'estimation de la longueur de la file d'attente [Negri 2014] ou même dans la navigation autonome.

- *Reconnaissance de l'objet.* Elle englobe toutes les applications de vidéosurveillance où l'identité des objets mobiles (piétons ou véhicules par exemple) est révélée par la détection d'éléments caractéristiques tels que : la reconnaissance de visage [Le 2014, Le 2012], la reconnaissance de plaque d'immatriculation, la classification des véhicules en mouvement [Duman 2013]. Aussi, les systèmes où les comportements des objets en mouvement peuvent être analysés, reconnus et interprétés [Amato 2011, Sanghyuk 2012].

- *Surveillance du trafic.* La surveillance automatique de la circulation joue un rôle essentiel dans le contrôle du trafic routier. Les systèmes avancés permettent de faciliter la gestion, la sécurisation et l'analyse de la circulation dans les réseaux routiers tels que : l'estimation des débits des flux de véhicules, le contrôle de vitesse des véhicules, le calcul de la densité de circulation sur l'autoroute [Tsung 2012, Xinfeng 2014], le contrôle du trafic aérien [Luo 2011] et maritime [Kruger 2012, Szpak 2011].

- *Détection d'événements anormaux.* Le but de ces systèmes est la surveillance des environnements, la détection d'événements anormaux et l'alerte dans certains cas, comme : la détection d'incendie [Lai 2007], la détection d'accident [Yun 2014], la détection de criminalité [Coppi 2011], la détection de tricherie [Dongbin 2014],....

2.4.4. Contribution à la classification suivant le niveau de l'application

Les systèmes de la vidéosurveillance intelligente peuvent offrir des résultats allant du bas niveau tels que la détection d'objet jusqu'à des niveaux très avancés tels que l'analyse comportementale des objets. Ces résultats sont très dépendants de l'application exigée du système. Suivant le niveau du résultat désiré, les systèmes de vidéosurveillance peuvent avoir différents niveaux de traitement. Hiérarchiquement, ils partent du niveau des pixels, en passant par les objets pour atteindre l'échelle des comportements. La question " Quel est le niveau de cette application de vidéosurveillance ? " peut avoir quatre réponses qui sont les quatre niveaux possibles d'une application de vidéosurveillance. Compte tenu de ces réponses, nous pouvons distinguer quatre tâches principales que le système peut exécuter en fonction de son niveau applicatif : la détection d'objet, le suivi d'objet, la classification et l'identification d'objet, l'analyse d'activités et du comportement d'objet. Ainsi, les systèmes de

vidéosurveillance peuvent être regroupés selon les tâches accomplies en quatre catégories, comme illustré dans la Figure 4.

- *Premier niveau.* Cette catégorie regroupe les applications de bas niveau pour les système de vidéosurveillance. La fonction de détection d'objets est suffisante pour ces simples applications. Dans la plupart des systèmes, les caméras utilisées sont supposées être statiques. Les applications de détection et/ou de comptage se basent principalement sur la fonction de détection d'objet sans avoir besoin d'atteindre des fonctions de niveaux supérieurs. Ces applications sont utilisées pour compter le nombre de personnes entrant et/ou sortant d'un bâtiment [Jun 2015, Yaning 2015], alerter quand il y a une activité dans une scène, estimer la longueur des files dans les magasins, surveiller les terminaux de bus ou les gares [Mukherjee 2011]. La détection et le comptage de véhicules est nécessaire pour calculer la congestion du trafic, estimer le débit des véhicules et garder la trace de véhicules qui utilisent un chemin particulier [Grzegorg 2014, Salvi 2014]. Les systèmes de détection et de comptage de personnes sont commercialisés aujourd'hui.

- *Deuxième niveau.* La tâche de suivi d'objet, précédée par sa détection, est utilisée par le niveau intermédiaire (deuxième niveau) des applications de vidéosurveillance. Elle vise à détecter la trajectoire d'un objet en mouvement dans la scène. Les applications appartenant à cette catégorie servent pour la surveillance de la congestion du trafic [Imran 2013], la détection des évènements anormaux [Sung 2014], la détection des disputes, la détection des objets abandonnés ou volés [Antonio 2013], la surveillance des personnes âgées [Bektuzun 2013], la détection de présence dans une zone interdite, le stationnement des voitures et l'arrêt brusque d'objets en mouvement [Maddalena 2013], ...

- *Troisième niveau.* Ces applications désignent les systèmes de classification et/ou d'identification. Les informations de suivi associées à divers attributs extraits (couleur, taille, etc.) sont les clés de l'analyse automatique des objets (type, l'identité, etc.) et leur recherche. Les objets détectés par un système de vidéosurveillance sont généralement classés en différentes catégories : homme, véhicule, animal, etc [Tuty 2014]. L'identification de l'objet détecté est en relation directe avec sa classe. Elle est faciale si l'objet est un être humain, ou se base sur l'analyse de la plaque d'immatriculation si l'objet est un véhicule.

– *Quatrième niveau.* L'analyse et la compréhension du comportement est considérée comme la tâche de plus haut niveau utilisée dans les applications de vidéosurveillance. C'est une étape essentielle dans laquelle les informations des fonctions de niveaux inférieurs sont combinées et interprétées par une description sémantique de haut niveau. La fonction d'analyse comportementale est utile dans les applications de détection d'événements exceptionnels, tels que la détection des suspects ou des personnes disparues [Chundi 2015], la détection de tricherie [Alayed 2013], la surveillance des personnes âgées [Héctor 2015], etc. Elle est également fiable pour l'analyse de la foule [Simone 2014] et l'analyse de la circulation [Brulin 2012].

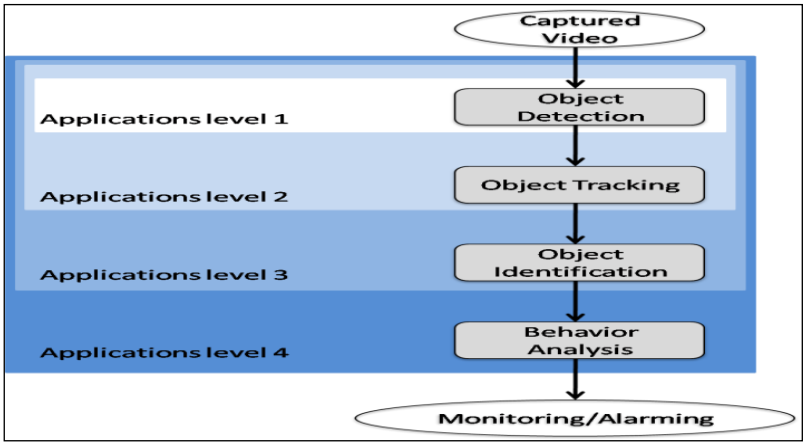

Figure 4. Les fonctions exécutées dans les systèmes de vidéosurveillance avec leurs niveaux d'applications correspondants

3. Analytique dans les systèmes de vidéosurveillance

Avec la multiplication des nombres de caméras de vidéosurveillance installées, le flux vidéo à archiver devient colossal ce qui dépasse les capacités des agents de surveillance. Pour traiter toutes ces informations, des logiciels intelligents ont été développés pour l'analyse automatique des scènes. D'où l'apparition de la notion de vidéosurveillance intelligente qui offre des systèmes permettant de détecter et suivre les objets et signaler les événements suspects. Toutes ces méthodes d'analyse intelligentes appartiennent à une technologie, encore récente, appelée l'analytique vidéo. Elle propose des solutions pour traiter la séquence vidéo enregistrée afin de n'en retenir que les données intéressantes pour la sécurité. Elle permet aussi d'améliorer les capacités de recherche dans les séquences archivées. La détection, le suivi, la classification et l'identification ainsi que l'analyse comportementale d'objets en mouvement sont, aujourd'hui, des techniques bien établies de l'analytique vidéo. Le flux de traitement est pratiquement toujours unidirectionnel. La Figure 5 illustre les quatre étapes de l'analytique vidéo.

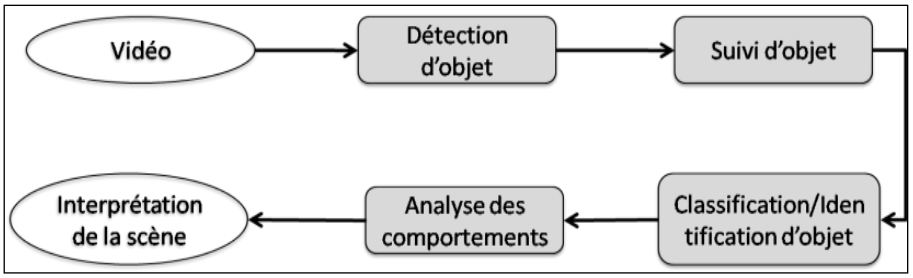

Figure 5. Les tâches d'analytique vidéo effectuées dans les systèmes de vidéosurveillance dans l'ordre croissant : de la tâche de bas niveau vers la tâche de haut niveau

3.1. *Détection d'objets en mouvement*

La détection d'objets est généralement à la base de tout système de vidéosurveillance intelligent. Elle permet de détecter les activités dans la scène surveillée, comme le mouvement, l'apparition ou la disparition d'un objet. La détection d'objets s'aligne avec la détection de mouvements puisque les régions en déplacement de la scène sont d'intérêt (l'avant-plan) et les parties statiques ne le sont pas (l'arrière-plan) [Ko 2008]. De nombreuses techniques de détection de mouvement se fondent sur la détection des changements. Cependant, la détection des changements dans une scène ne peut pas cibler nécessairement le mouvement d'objets, mais elle peut mettre en évidence une modulation de l'image. Pour segmenter les objets en mouvement, nous devons réussir à discriminer entre les pixels correspondant à des mouvements cohérents et ceux causés par les changements environnementaux. Les environnements complexes peuvent présenter un problème majeur en raison de nombreuses variantes (changements d'éclairage, mouvements inutiles, arrière-plans encombrés). Plusieurs techniques de segmentation de mouvement sont couramment utilisées dans la littérature telles que :

3.1.1. Différences temporelles

Une première classe de méthodes de détection de mouvement se base sur une différenciation temporelle, entre les images. Elles ne demandent pas de modèles d'arrière-plan et extraient les régions en mouvement par l'analyse de la variation temporelle de l'intensité lumineuse des pixels. Celles-ci sont très rapides et s'adaptent aux environnements dynamiques. L'idée de base pour extraire les zones en déplacement consiste à calculer la différence absolue Δ_t entre n trames consécutives.

$$\Delta_t(x, y) = |I_t(x, y) - I_{t-n}(x, y)| \quad (1)$$

Avec, $I_t(x, y)$ l'intensité lumineuse du pixel de coordonnées (x, y) de la t ième image et n varie entre 1 et 5. Un seuil fixé τ permet de distinguer entre les pixels appartenant à l'avant-

plan et ceux de l'arrière-plan. L'image des zones d'avant-plan $F(x, y)$ est extraite par seuillage.

$$\boldsymbol{F_t(x,y)} = \begin{cases} \mathbf{1} & \boldsymbol{si}\ \Delta_t(x,y) \geq \tau \\ \mathbf{0} & \boldsymbol{sinon} \end{cases} \quad (2)$$

Le choix de la méthode de seuillage influence les résultats de la segmentation. Un simple seuillage ne permet pas souvent d'extraire tous les pixels concernés des objets surtout si ceux-ci se déplacent lentement. Des trous apparaissent à l'intérieur des entités mobiles, ce qui nécessite des opérations morphologiques pour lisser ces défauts de segmentation. Pour résoudre ces problèmes, certains auteurs, comme [Kameda 1996], proposent une variante de cette méthode " La double différence ". Cette approche calcule la différence entre les images t et $t-1$ et entre les images $t-1$ et $t-2$, en les combinant avec un ET logique. Un autre travail [Collins 2000], dans le cadre du projet VSAM, propose un algorithme qui exploite la différence entre les images t et $t-1$ et la différence entre t et $t-2$.

3.1.2. Flot optique

Le flot optique décrit le taux directionnel et temporel de pixels dans deux images successives d'une séquence vidéo [Aslani 2013]. Un vecteur de vitesse à deux dimensions portant des informations sur la direction et la vitesse du mouvement est attribué à chaque pixel de l'image. Pour avoir un calcul plus simple et plus rapide, le monde réel en trois dimensions (3D + temps) est transféré en (2D + temps). Ensuite, l'image est décrite à l'aide d'une fonction dynamique de luminosité $I\ (x, y, t)$ dépendante des coordonnées du pixel et du temps. En supposant que dans le voisinage d'un pixel déplacé, le changement d'intensité de luminosité ne se produit pas le long du champ de mouvement, l'expression suivante est déduite :

$$\boldsymbol{I\ (x,y,t) = I\ (x+\delta x, y+\delta y, t+\delta t)} \quad (3)$$

En appliquant les séries de Taylor sur la partie droite de (3), on obtient :

$$\boldsymbol{I\ (x+\delta x, y+\delta y, t+\delta t) = I\ (x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + H.O.T} \quad (4)$$

A partir de (3) et (4), en négligeant les termes d'ordres supérieurs $(H.O.T)$ et après modifications, on obtient :

$$\boldsymbol{I_x.v_x + I_y.v_y = -I_t} \quad (5)$$

Qui se traduit dans la représentation formelle du vecteur par :

$$\boldsymbol{\nabla I.\vec{v} = -I_t} \quad (6)$$

ou ∇I est le gradient spatial de l'intensité de luminosité, $\vec{v}$ le flot optique (vecteur de vitesse) du pixel et I_t la dérivée dans le temps de l'intensité de luminosité.

L'équation (6), à deux inconnues, est la plus importante pour le calcul du flot optique. Elle est appelée l'équation de contrainte de mouvement 2D ou le gradient de contrainte. L'estimation du flot optique nécessite beaucoup de calcul. A présent, il existe plusieurs méthodes pour sa résolution. Toutes ces méthodes partent de l'équation (3) qui suppose la conservation d'intensité de luminosité. La détermination du flot optique est résolue par le calcul des dérivées partielles du signal de l'image. Deux méthodes sont les plus utilisées, à savoir : Lucas-Kanade [Lucas 1981] et Horn-Schunck [Horn 1981]. Ces deux méthodes supposent que la luminosité ne change pas au fil du temps. La méthode de Lucas-Kanade est une méthode disséminée et locale, tandis que la méthode Horn-Schunck est dense et globale. Cette dernière suppose que le champ d'écoulement est globalement lisse (les vitesses voisines sont presque identiques). Alors que la méthode de Lucas-Kanade suppose que la vitesse est localement constante, et que les points voisins ont des déplacements semblables. La méthode de Lucas-Kanade produit moins de bruit par rapport à la méthode de Horn-Schunck.

Le flot optique est utilisé pour détecter les objets en mouvement de façon indépendante en présence des mouvements de la caméra. Cependant, la plupart de ses méthodes exigent des calculs complexes difficiles à exécuter en temps réel. De plus, le flot optique est sensible au bruit de l'image.

3.1.3. Soustraction de l'arrière-plan

La soustraction de l'arrière-plan se compose de deux grandes étapes : (1) Modélisation de l'arrière-plan; (2) Segmentation de mouvement. La modélisation de fond est la représentation de la scène sans les objets en mouvement et elle doit être mise à jour régulièrement. La segmentation de mouvement vise à détecter les régions correspondantes à des objets mobiles (personnes, véhicules, ...). Les images de la vidéo sont comparées au modèle de fond et les différences sont marquées comme des objets en mouvement. Dans ce qui suit, nous classifions les principales méthodes de soustraction de fond en deux catégories : les méthodes récursives et les méthodes non récursives. La différence entre les deux est que les premières utilisent une valeur unique de l'arrière-plan. Les autres utilisent un buffer pour représenter l'arrière-plan.

3.1.3.1. Techniques non récursives

Une technique non-récursive utilise une approche de fenêtre glissante pour l'estimation de fond. Il stocke dans un buffer les L dernières trames de la séquence vidéo, et estime l'image de l'arrière-plan en se basant sur la variation temporelle de chaque pixel au sein du buffer. Les techniques non récursives sont très adaptatives, car elles ne dépendent que des images stockées dans le buffer. D'autre part, en cas de lents mouvements dans la séquence, un buffer de taille significative est nécessaire. Ce problème peut être partiellement atténué en stockant les images à une fréquence temporelle plus faible. Certaines techniques non récursives, qui sont couramment utilisées, sont décrites ci-dessous [Cheung 2004, Setitra 2014] :

- *Différence entre images.* C'est la plus simple technique de modélisation de fond. Elle utilise la trame à l'instant $t-1$ comme modèle de fond pour la trame à l'instant t. Ceci est un avantage qui rend les consommations en termes de calcul et de temps faibles. L'aspect de fond multimodal ne peut pas être maintenue puisque l'historique des pixels n'est pas connue. Ainsi, cette technique ne peut pas éviter la détection d'objets non essentiels tels que des branches d'arbres en mouvement.

- *Filtre médian.* Dans le filtre temporel médian [Lo 2001], le modèle de fond est réalisé en calculant la valeur de la médiane d'un buffer des L dernières images. Ensuite, chaque nouveau pixel est comparé au modèle. Si la différence respecte un certain seuil, il est considéré comme fond, sinon il est considéré comme avant-plan. La mise à jour du modèle de l'arrière-plan se fait en ajoutant la valeur actuelle du pixel dans la mémoire du buffer tant que sa taille le permet. Le procédé a été amélioré par Cucchiara et al. dans [Grana 2001] et [Calderara 2006]. Dans [Grana 2001] [Cheung 2005], les auteurs utilisent la médoïde à la place du filtre médian pour les images en couleurs. Bien que ces deux filtrages, médian et médoïde, ont montré leurs performances en dépit de leur simplicité, ils souffrent encore des hautes exigences en mémoire dû à la taille du buffer nécessaire pour la modélisation. La complexité de calcul de la méthode du filtre médian est $O\ (L\ log\ L)$ pour chaque pixel.

- *Filtre prédictif linéaire.* Cette méthode estime l'arrière-plan en appliquant un modèle de filtre prédictif linéaire au buffer. A chaque trame, les coefficients du filtre sont estimés. L'estimation est basée sur le calcul des covariances des échantillons. Cette méthode présente l'inconvénient d'être lourde en calcul et en consommation de

mémoire, ce qui la rend difficile à appliquer en temps réel. Comme exemple de filtre prédictif utilisé, le filtre de Wiener [Toyama 1999]. La prédiction x_t de la valeur du pixel à l'instant t est donnée par :

$$x_t = -\sum_{k=1}^{p} a_k \, x_{t-k} \qquad (7)$$

avec a_k les coefficients de prédiction du filtre qui sont déterminés à partir de la covariance des valeurs de x_t. Le filtre utilise les p échantillons les plus récents pour effectuer la prédiction. La décision de classification utilise l'erreur de prédiction e, définie par :

$$E[e_t^2] = E[s_t^2] + \sum_{k=1}^{p} a_k \, E[s_t^2 - k] \qquad (8)$$

Pour chaque pixel, cette erreur est évaluée et si un pixel s'écarte de plus de $4\sqrt{E[e_t^2]}$, il est considéré comme étant en mouvement.

– *Modèle non-paramétrique.* Le modèle non-paramétrique [Elgammal 1999], effectue la soustraction de fond en utilisant toute l'histoire du pixel. L'estimation du modèle de fond se fait en utilisant une fonction noyau. Dans [71] la fonction noyau utilisée est une gaussienne. Toute l'historique du pixel $I_{t-L}, I_{t-L+1}, \dots, I_{t-1}$ est utilisée pour former une estimation non paramétrique de la fonction de densité du pixel $f(I_t = u)$:

$$f(I_t = u) = \frac{1}{L}\sum_{i=t-L}^{t-1} K(u - I_i) \qquad (9)$$

$K(.)$ est l'estimateur du noyau qui a été choisi pour être gaussien. Le pixel courant I_t est déclaré comme avant-plan si $f(I_t)$ est inférieure à un certain seuil prédéfini. Cette méthode présente l'avantage de traiter la soustraction multimodale de l'arrière-plan mais elle consomme beaucoup en terme de temps et de mémoire.

3.1.3.2. Techniques récursives

Les techniques récursives ne maintiennent pas un buffer pour l'estimation du fond. Elles mettent à jour de manière récursive un seul modèle de fond en fonction de chaque trame. Par conséquent, les trames anciennes peuvent avoir un effet sur le modèle d'arrière-plan actuel. Par rapport aux techniques non récursives, celles récursives nécessitent moins de mémoire, mais une erreur dans le modèle d'arrière-plan peut persister pendant une longue période de temps. Certaines techniques récursives sont décrites ci-dessous :

– *Filtre médian approximatif.* En raison du succès du filtre médian non-récursif, les auteurs de [Farlane 1995] propose un filtre récursif simple pour la soustraction du fond. L'arrière-plan est d'abord initialisé à la première image, puis, pour chaque

nouvelle observation, le pixel est comparé à l'arrière-plan. La mise à jour du fond est faite de la manière suivante : si le pixel de l'avant-plan est supérieur à celui de l'arrière-plan, ce dernier est incrémenté; s'il est inférieur, le fond est diminué; s'ils sont égaux, le fond reste le même. Cette méthode est simple, robuste aux bruits et faible en taux de calcul. D'autre part, elle ne garde pas l'historique des pixels et ne modélise pas leurs variances.

– *Filtre Kalman.* Le filtre Kalman est une technique récursive largement utilisée. Différentes versions ont été proposées pour la modélisation de fond. La version la plus simple utilise seulement l'intensité de luminance [Wren 1997, Heikkila 1999, Halevy 1999, Boult 1999]. Kalman et al. utilisent à la fois l'intensité et sa dérivée temporelle [Karmann 1990], tandis que Koller et al. utilisent l'intensité et ses dérivées spatiales [Koller 1993]. Le schéma proposé dans [Karmann 1990] décrit l'état interne du système par l'intensité de son arrière-plan B_t et sa dérivé temporelle B'_t, qui sont mises à jour d'une manière récursive comme suit :

$$\begin{bmatrix} B_t \\ B'_t \end{bmatrix} = A.\begin{bmatrix} B_{t-1} \\ B'_{t-1} \end{bmatrix} + k_t.\left(I_t - H.A.\begin{bmatrix} B_{t-1} \\ B'_{t-1} \end{bmatrix}\right) \quad (10)$$

Avec I_t l'intensité de luminance du pixel à l'instant t, A est la matrice décrivant la dynamique de l'arrière-plan et H la matrice de mesure. Leurs valeurs particulières utilisées dans [Karmann 1990] sont les suivantes :

$$A = \begin{bmatrix} 1 & 0.7 \\ 0 & 0.7 \end{bmatrix}, H = [1 \quad 0] \quad (11)$$

La matrice de gain de Kalman K_t varie entre un taux d'adaptation lent α_1 et un taux d'adaptation rapide α_2 selon si I_t est un pixel de l'avant-plan ou non :

$$K_t = \begin{cases} \begin{bmatrix} \alpha_1 \\ \alpha_1 \end{bmatrix}, \text{ si } I_{t-1} \text{est de l'avant} - \text{plan} \\ \begin{bmatrix} \alpha_2 \\ \alpha_2 \end{bmatrix}, \text{ sinon} \end{cases} \quad (12)$$

– *Mélange de gaussiennes.* Contrairement au filtre Kalman qui permet de suivre l'évolution d'une seule gaussienne, la méthode de mélange de gaussiennes [Stauffer 1999] suit multiple distributions simultanément. Elle maintient une fonction de densité pour chaque pixel. Ainsi, elle est capable de manipuler des distributions de fond multimodales. D'autre part, puisque le mélange de gaussiennes est paramétrique, les paramètres du modèle peuvent être mis à jour de manière adaptative sans garder un grand buffer des images. Les valeurs d'un pixel particulier sont modélisées par $k = 3$

ou $k = 5$ distributions gaussiennes. Pour chaque i-ème gaussienne à l'instant t, $\omega_{i,t}$ est le poids, $\mu_{i,t}$ est la valeur moyenne et $\Sigma_{i,t}$ est la covariance. La probabilité P de l'observation de la valeur de pixel courant X_t est :

$$P(X_t) = \sum_{i=1}^{k}(\omega_{i.t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t})) \qquad (13)$$

$\eta(X_t, \mu_{i,t}, \Sigma_{i,t})$ est une fonction de densité de probabilité gaussienne. La covariance $\Sigma_{i,t}$ est supposée être diagonale, avec $\sigma_{i,t}^2$ les éléments de sa diagonale. Chaque nouvelle valeur de pixel, est vérifiée contre les k distributions gaussiennes existantes jusqu'à ce qu'une correspondance est trouvée. Les poids des k distributions à l'instant t sont ajustés comme suit :

$$\omega_{k,t} = (1 - \alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \qquad (14)$$

α est le taux d'apprentissage et $M_{k,t}$ est égal à 1 pour le modèle qui correspond et égal à 0 pour les autres modèles. Les paramètres de la distribution qui correspondent à la nouvelle observation sont mis à jour comme suit :

$$\mu_{i,t} = (1 - \rho)\mu_{i,t-1} + \rho.X_t \qquad (15)$$

$$\sigma_{i,t}^2 = (1 - \rho)\,\sigma_{i,t-1}^2 + \rho\,(X_t - \mu_{i,t})^T\,(X_t\text{-}\mu_{i,t}) \qquad (16)$$

avec $\rho = \alpha.\eta(X_t|\mu_k, \sigma_k)$, $\sigma_{i,t-1}^2$ est la dernière variance et $(X_t - \mu_{i,t})^T\,(X_t\text{-}\mu_{i,t})$ est la distance entre le nouveau pixel et la moyenne mise à jour. Si aucune distribution ne correspond à la valeur du pixel courant, la distribution ayant la plus faible probabilité sort. Une nouvelle distribution avec la valeur actuelle de sa valeur moyenne, une variance initialement élevée, et un faible poids entre. Les paramètres μ et σ des distributions non correspondantes restent les mêmes. Une fois les gaussiennes sont mises à jour, elles sont triées selon la valeur de ω/σ. Ensuite, les B premières distributions sont choisies en tant que le modèle d'arrière-plan, où :

$$B = \text{argmin}_k\left(\sum_{i=1}^{k}\omega_i > T_{bg}\right) \qquad (17)$$

T_{bg} est une mesure de la partie minimale des données qui doit être représentée par le fond. Le mélange de gaussiennes est robuste face aux changements d'éclairage ralentis, mouvements périodiques dans un milieu encombré, objets ayant des mouvements lents, changements de scène à long terme, et aux bruits de la caméra. Plusieurs améliorations ont été proposées pour le mélange des gaussiennes comme : Zivkovic et al. qui proposent dans [Zivkovic 2004, Zivkovic 2006] un mélange des gaussiennes adaptatif. La méthode adapte automatiquement le nombre de gaussiennes utilisées pour modéliser un pixel donné. Cette extension permet de réduire les besoins

en mémoire, augmenter l'efficacité de son calcul, et améliorer les performances lorsque le fond est très multimodal.

3.2. *Suivi d'objets en mouvement*

Après la détection des objets en mouvement, leurs déplacements sont suivis tout au long de la séquence vidéo. Le suivi est l'estimation de la trajectoire d'un objet dans le plan image comme il se déplace dans la scène. Cette tâche requiert localiser chaque objet à partir d'une image à l'autre. Le suivi peut être fait en 2D, à partir d'une seule caméra, ou 3D, en combinant deux vues ayant une relation géométrique connue. De nombreuses techniques de suivi prédisent la position de l'objet dans une trame à partir de ses déplacements observés dans les images précédentes. Chaque objet détecté doit être associé à son correspondant dans la trame suivante pour mettre à jour sa trajectoire, sinon une nouvelle trajectoire est créée. Le suivi de ces objets peut être difficile en raison de la complexité de leurs formes, leur nature non-rigide, leurs mouvements, des occlusions partielles ou complètes, des changements d'éclairage de la scène, etc. Ceux-ci peuvent être simplifiés par une simple supposition comme les mouvements lisses, et la connaissance préalable du nombre, de la taille, de la forme et de l'apparition des objets. Le suivi permet d'extraire d'autres caractéristiques : la trajectoire, la vitesse, la direction du mouvement, la position à un moment précis. Il existe différentes classifications disponibles pour les méthodes de suivi d'objets, comme [Meng 2015, Yilmaz 2006, Moeslund 2006, Hu 2004]. Les classifications les plus populaires sont celle de [Hu 2004] qui classe les algorithmes de suivi d'objets en quatre catégories : algorithmes basés sur les régions, algorithmes basés sur les contours actifs, algorithmes basés sur les caractéristiques et algorithmes basés sur les modèles, et celle de [Yilmaz 2006] où les auteurs classent les algorithmes en trois catégories : suivi de points, suivi à noyaux et le suivi de silhouette. Dans cette section, nous reprenons la classification proposée par [Meng 2015] qui est une classification plus récente et plus exhaustive. Selon les différentes méthodologies des algorithmes de suivi d'objets, Li et al. classent les méthodes de suivi existantes en quatre catégories : suivi basé sur le matching (appariement), suivi basé sur le filtrage, suivi basé sur la classe et suivi basé sur la fusion.

3.2.1. Approches basées sur le matching

Les algorithmes de suivi basés sur le matching sont établis pour chercher une correspondance au modèle d'objet avant de le suivre. En fonction du résultat de matching, le

suivi est assuré tout au long de la séquence vidéo. Avec différentes descriptions des attributs de l'objet, nous pouvons diviser les algorithmes de cette catégorie en quatre méthodes de base.

- *Suivi basé sur la région.* L'idée fondamentale du suivi basé sur la région est : la région initiale de l'objet dans l'image forme le modèle de l'objet, la correspondance au modèle est cherchée dans tous les emplacements possibles des images candidates, la position ayant le degré de correspondance le plus élevé est jugée la meilleure correspondance, et la région identifiée par cette position est la nouvelle région de l'objet [Hager 1998]. Le plus commun critère de mesure de corrélation est la somme des différences carrées SSD (Sum of Square Difference). Lucas et al. [Lucas 1981] présentent une méthode utilisant le gradient spatial de l'image en niveau de gris pour trouver la meilleure région correspondante. Ils utilisent les valeurs de gradient de chaque point dans la région de l'objet pour mettre à jour la région candidate de l'objet. Cette méthode peut être appliquée au suivi des mouvements affines ou des objets non rigides. Il est possible d'utiliser un modèle fixe pour une courte période de suivi, mais il est déconseillé pour une longue période de suivi en raison des changements d'apparence de l'objet. Si les modèles d'objets peuvent changer avec les apparitions d'objets, la fiabilité des algorithmes de suivi sera améliorée. Jepson et al. [Jepson 2003] proposent un modèle adaptatif basé sur les caractéristiques de la texture, ce qui assure une robustesse face aux occlusions. Zoidi et al. [Zoidi 2013] exploitent la soustraction de fond pour détecter et mettre à jour le modèle par la similitude des couleurs de l'histogramme. Le système proposé assure un suivi réussi sous des variations d'échelle, des rotations et des occlusions partielles. L'algorithme MeanShift [Comaniciu 2002] est un représentant typique des algorithmes de suivi basés sur la région. Un grand nombre d'algorithmes améliorés [Nouar 2006,Leitcher 2012] sur la base MeanShift sont proposés. Les algorithmes de suivi basés sur la région utilise des informations globales sur l'objet, telles que les informations couleurs, les caractéristiques de la texture, etc. Pour cela, d'une part ils ont une grande crédibilité, d'autre part une faible déformation des objets n'a aucune incidence sur les performances de suivi. Malheureusement, ils consomment beaucoup de temps en calcul lorsque les régions de recherche sont larges. En plus, ils ne sont pas exacts lorsque les objets ont de grandes déformations ou de fortes occlusions. Au cours des dernières années, les algorithmes de suivi de la région se concentrent sur la façon de

s'adapter avec les variations des modèles comme les travaux de [Cannons 2014, He 2014]. Les variations sont causées par les diverses poses des objets en mouvement.

- *Suivi basé sur les caractéristiques.* L'approche fondamentale des algorithmes de suivi basés sur les caractéristiques [Sinha 2011] est : identifier l'objet par ses caractéristiques, et chercher son correspondant dans la séquence vidéo en se basant sur les caractéristiques de l'image. Cet algorithme a généralement deux principales étapes qui sont l'extraction des caractéristiques et leur appariement. La première étape vise à extraire les principales caractéristiques comme sommet, centre de gravité, etc. La deuxième étape est de trouver l'objet le plus similaire dans la trame suivante selon un critère de correspondance. Ainsi, la position de l'objet est déterminée dans toute la séquence vidéo. Les premiers travaux identifient les points caractéristiques correspondants entre les trames adjacentes. La méthode courante consiste à supposer certaines conditions sur le mouvement des points caractéristiques. Sethi et al. [Sethi 1987] supposent que le mouvement des points est lisse pour assurer le suivi. Tissainayagam et al. [Tissainayagam 2005] proposent un algorithme de suivi qui cherche les points des contours des objets ayant un maximum local, et les définit comme les points clés pour le suivi. La performance de cette technique est bonne pour les objets géométriques simples. Cependant, elle s'atténue pour les objets complexes due à la difficulté d'extraire les points de coins stables. Les algorithmes de suivi basés sur les caractéristiques ne peuvent pas traiter efficacement les occlusions et les chevauchements. Récemment, Li et al. [Li 2012] présente une nouvelle méthode pour savoir comment sélectionner les caractéristiques pertinentes, et améliorer les performances en temps réel de ces algorithmes.

- *Suivi basé sur le modèle déformable.* Le principe fondamental de ces algorithmes [Zhong 2000] est : utiliser la surface ou la courbe du contour de l'objet en mouvement qui a de bonnes propriétés d'élasticité et de déformation comme le contour délimitant de l'objet, et mettre à jour ce contour pour correspondre à l'objet. Kass et al [Kass 1988], proposent le modèle des contours actifs qui est le modèle déformable le plus couramment utilisé dans le suivi. Les Snakes sont des modèles de contours actifs qui s'arrêtent sur les bords et localisent les objets avec précision. De nombreux algorithmes améliorés à base de Snakes ont été proposés [Ronfard 1994, Brigger 2000, Paragios 2002, Xue 2002]. L'utilisation des contours actifs est réussie pour le suivi des

mouvements des objets rigides et non rigides [Derrode 2006, Fang 2011]. Dans les situations d'interférences de bruits, des occlusions, et des bords flous, il est très difficile d'obtenir des contour précis. Les informations sur la couleur [Li 2004], la texture [Houhou 2008], et la forme [Charpiat 2007] sont utilisées pour contraindre les contours à obtenir des bords exacts. En revanche, les algorithmes de suivi basés sur des modèles déformables décrivent simplement les objets, réduisent la complexité de calcul et ont une grande robustesse dans les situations d'occlusions partielles. Cependant, il est difficile de démarrer automatiquement le suivi avec ces algorithmes, car leur initialisation est très sensible.

- *Suivi basé sur le modèle géométrique.* Ces algorithmes se basent sur : l'établissement d'un modèle géométrique de l'objet en fonction des connaissances a priori et le suivi en faisant correspondre le modèle de la région candidate et le modèle de l'objet [Dahlkamp 2007]. Les modèles peuvent être divisés en : modèle hiérarchique [Karaulova 2002], modèle 2D [Wu 2004] et modèle 3D [Gavrila 1996]. Yang et al., [Yang 2001] présentent un algorithme de localisation de véhicules basé sur un modèle 3D, qui peut de manière efficace et robuste déterminer les positions des véhicules dans les scènes de circulation par des caméras calibrées. La méthode utilise les points de bords dans les images comme des caractéristiques, et mesure le degré de correspondance entre ces points et le modèle prévu. Les algorithmes de suivi basés sur des modèles ne sont pas facilement affectés par les perspectives d'observation, de sorte qu'ils sont intrinsèquement robustes à différents mouvements. Inéluctablement, ces algorithmes ont des insuffisances telles que la nécessité de construire les modèles, les coûts de calcul élevés, les mécanismes complexes de mise à jour des modèles, et les faibles performances en temps réel.

3.2.2. Approches basées sur le filtrage

Les algorithmes de suivi à base de filtrage considèrent que les problèmes de suivi sont des problèmes d'estimation d'états [Williams 2005]. Les états d'un objet peuvent inclure toutes les caractéristiques des mouvements. La clé de suivi est de savoir comment déduire à posteriori la densité de probabilité des états. Afin de la calculer, deux modèles sont introduits comme suit :

Modèle d'état : $$S_k = f_k(S_{k-1}, W_k) \qquad (18)$$

Modèle d'observation : $Z_k = h\,(S_k, V_k)$ (19)

Le modèle d'état est utilisé pour la description de l'évolution du système, et le modèle d'observation est utilisé pour la description de la relation entre l'observation et l'état. La fonction de densité de probabilité postérieure comprend toutes les informations statistiques sur les états de l'objet qui peuvent être obtenus durant le processus de suivi. Elle est donc une solution complète au problème d'estimation. Le filtre de Kalman [Kalman 1960] est une méthode efficace pour estimer les états. Il les prédit à partir du modèle d'état, et estime la fonction de densité de probabilité a posteriori à partir du modèle d'observation. Le filtre à particules [Gordon 1993] est un filtre séquentiel, qui permet de résoudre le problème de l'estimation sous la condition de non-linéairité et non-Gaussienne. Un grand nombre d'études montrent que dans les environnements complexes, le suivi avec le filtre à particules a de meilleures performances que le filtre de Kalman. Cependant, le filtre à particules souffre des problèmes de précision de suivi lors des occlusions.

3.2.3. Approches basées sur la classe

Au cours des dernières années, certains chercheurs [Andriluka 2008, Ross 2008, Babenko 2011] considèrent le problème de suivi comme un problème de classification d'avant-plan et d'arrière-plan. Ils déduisent précisément les positions des objets en construisant des classificateurs pour classer les zones de localisation. Avidan [Avidan 2004] combine les méthodes de flux optique et de SVM (Support Vector Machine), et les utilise pour le suivi des véhicules. Kalal et al. [Kalal 2012] développent une nouvelle méthode d'apprentissage pour estimer les erreurs de suivi et mettre à jour le système pour les éviter dans l'image suivante. Cette méthode montre une amélioration significative par rapport aux approches de l'état de l'art durant le suivi à long terme. Ces algorithmes ont une haute précision de suivi. Malheureusement, ils ont deux inconvénients : d'une part, la construction des classificateurs a besoin de plusieurs échantillons positifs et négatifs d'où la difficulté d'étudier et de sélectionner ces échantillons; d'autre part, la nécessité de chercher les objets dans une large zone.

3.2.4. Approches basées sur la fusion

Dans les applications pratiques, les algorithmes de suivi basés sur la fusion sont proposés pour atteindre de bons résultats de suivi. Ces approches combinent souvent une variété d'algorithmes ou différentes sources d'information pour améliorer la précision des résultats. Ces méthodes permettent d'utiliser pleinement les avantages complémentaires de

différentes techniques pour obtenir un suivi de haute qualité. Ces algorithmes peuvent être divisés en trois types :

Le premier type de méthodes est basé sur la fusion multi-caractéristiques, qui est l'approche la plus commune. On ne peut pas avoir un suivi stable à long terme en utilisant une caractéristique unique, en raison de la complexité des scénarios ou des objets. Par conséquent, de nombreux chercheurs utilisent des méthodes de fusion d'informations pour améliorer les performances de suivi. Ils donnent des poids différents suivant la capacité de la caractéristique de décrire les objets en mouvement. Zhou et al. [Zhou 2006] présente une méthode qui intègre la position spatiale, la forme et les informations de couleurs pour améliorer les performances

La seconde catégorie de méthodes est basée sur la fusion multi-modèles. Cette approche intègre des modèles d'objets dans différents instants de la séquences vidéo ou combine des modèles d'objets avec différents angles de plusieurs caméras. Cette approche améliore la robustesse du suivi, en raison de l'adaptation aux changements de l'objet et la détermination des caractéristiques efficaces. Khan et al., [Khan 2009] proposent une approche multi-vues pour résoudre les problèmes de suivi dans les scènes encombrées.

Le troisième type de méthodes est basé sur la fusion multi-algorithmes. Différents algorithmes de suivi d'objets ont leurs propres avantages pour certaines scènes. La méthode permet d'utiliser pleinement les performances des divers algorithmes en intégrant les méthodes appropriées, afin qu'elle puisse surmonter les faiblesses d'un algorithme individuel. Shan et al. [Shan 2007] propose une intégration de l'algorithme de MeanShift au filtre à particules. Les algorithmes de fusion peuvent avoir plus de précision des résultats. Malheureusement, le temps de calcul est plus long.

3.3. *Classification d'objets en mouvement*

La classification est une tâche de reconnaissance d'objets. Pour les suivre et analyser leurs comportements, il est essentiel de les classer correctement. Les objets détectés peuvent être généralement classés en véhicules, animaux, humains, arbres se balançant et d'autres objets en mouvement [Tuty 2014]. En général, le système reconnaît la nature d'une entité détectée à partir des attributs de sa forme et/ou des propriétés de son mouvement [Ko 2008]. Les approches de classification sont basées sur le mouvement, la forme, la couleur et la texture.

3.3.1. Approches basées sur la forme

La classification basée sur la forme s'intéresse purement à la géométrie de l'objet. Selon la géométrie des régions extraites, comme les boîtes englobantes, les contours externes, les objets peuvent être classés. Les auteurs de [Hota 2007] explorent l'étude de diverses caractéristiques des formes avec précision. Tsai et al. [Tsai 2006] présentent une méthode pour suivre les humains dans les scènes encombrées, en faisant recours aux modèles de formes humaines, en plus des modèles de caméra. Selon [Lee 2011], la classification basée sur la forme a une précision raisonnable. Son temps de calcul est considéré faible par rapport à d'autres méthodes de classification.

3.3.2. Approches basées sur le mouvement

L'approche basée sur le mouvement offre une méthode robuste pour la classification [Javed 2002]. Elle ne nécessite pas de modèles de forme prédéfinis, mais elle a du mal à identifier un objet non mobile [Parekh 2014]. Bien que la classification basée sur le mouvement a une précision modérée, son calcul est peu coûteux. Les déplacements non rigides des objets articulés présentent une propriété périodique très intéressante pour leur classification. Le flot optique est également très utile : le flux résiduel peut être utilisé pour analyser la rigidité et la périodicité des entités en mouvement. Il est prévu que les objets rigides présentent peu de flux résiduel que ceux non rigides [Hitesh 2013]. Johnsen et al. [Johnsen 2009] proposent un système de suivi et de classification qui a montré de bons résultats sur multiples objets dans des conditions variées de luminosité et d'occlusions.

3.3.3. Approches basées sur la couleur

Contrairement à beaucoup d'autres caractéristiques de l'image (par exemple, la forme), la couleur est relativement constante durant les changements de point de vue et elle est facile à acquérir. La représentation des caractéristiques de la couleur est le moyen le plus efficace pour révéler la similitude des images couleurs. Dans les systèmes de recherche d'images par le contenu [Sergyn 2007], les recherches les plus simples et les plus efficaces sont celles basées sur la couleur. Les auteurs de [Mahalingam 2010] proposent un système de suivi d'objets en mouvement basé sur la segmentation des images couleurs et l'histogramme des couleurs. Selon [Parekh 2014], la précision et le temps de calcul sont élevés pour la classification basée sur la couleur.

3.3.4. Approches basées sur la texture

L'affectation d'une image à une classe de texture connue est un objectif important de la classification basée sur la texture. Avec l'existence de plusieurs classificateurs, la tâche principale est l'extraction des caractéristiques pertinentes de l'image texturée. Ces approches se composent de deux phases : la phase d'apprentissage et la phase de reconnaissance. Les méthodes basées sur la texture telles que les histogrammes de gradient orienté HOG (Histogram of Oriented Gradient) utilisent les caractéristiques dimensionnelles basées sur les contours [Dalal 2005]. Conformément à [Parekh 2014], ces méthodes donnent une meilleure précision mais avec un temps de calcul supplémentaire.

Après avoir déterminé la classe à laquelle appartient un objet, son identité doit être révélée. Dans les systèmes de surveillance pour le contrôle d'accès ou la recherche de suspects [Wang 2012], en plus de la classification, l'identité de l'objet doit aussi être dévoilée par exemple par la reconnaissance faciale de l'individu ou la lecture de la plaque d'immatriculation de la voiture. Beaucoup de recherches ont été investies au cours des dernières années dans ces deux applications spécialisées. La reconnaissance faciale est parmi les principaux outils utilisés pour l'identification biométrique des personnes en vidéo [Ibrahim 2012] car elle permet une identification plus précise. Toutefois, la reconnaissance de visage dans un environnement non contrôlé reste un problème qui n'est pas encore résolu de manière satisfaisante. La lecture des plaques d'immatriculation dans les systèmes de surveillance vidéo est une application difficile [Chen 2013]. Elle exige avoir une image à haute résolution. L'analyse de l'image est confrontée à nombreuses interférences environnementales. Pour maximiser l'efficacité, la reconnaissance de plaque est le plus souvent réalisée par des systèmes spécialisés avec des caméras bien positionnées et une qualité d'éclairage adéquate.

3.4. ***Analyse comportementale d'objets en mouvement***

L'analyse des comportements est la tâche de plus haut niveau utilisée par les systèmes de la vidéosurveillance intelligents. Les informations collectées par les étapes précédentes sont interprétées par une description sémantique pour décrire les comportements et les interactions des objets de la scène avec un langage naturel. L'analyse sémantique est souvent très dépendante du contexte de l'application. Les techniques les plus utilisées pour modéliser les comportements détectés sont : les modèles de Markov cachés, les réseaux de neurones, les réseaux Bayésiens [Weiming 2004], etc. Tout d'abord, les informations visuelles des objets en

mouvement de la scène sont extraites et décrites avec une méthode appropriée, ensuite ces informations sont étudiées pour reconnaître et comprendre le comportement. De nombreuses caractéristiques ont été proposées pour décrire les activités humaines en se basant sur trois algorithmes principaux [Bobick 2001] :

- *Algorithmes basés sur les modèles 3D.* La technique la plus courante pour atteindre les informations 3D d'un mouvement est de récupérer la pose de la personne ou de l'objet à chaque instant en utilisant un modèle 3D. Le modèle est construit en essayant de minimiser une mesure résiduelle entre le modèle projeté et les contours de l'objet. Cela nécessite généralement une forte segmentation avant-plan/arrière-plan. Comme exemple, Campbell et Bobick [Campbell 1995] qui calculent les informations 3D des positions des membres du corps humain. Leur système exploite les redondances qui existent pour des actions particulières et effectue la reconnaissance en utilisant uniquement l'information qui varie entre ces actions. Cette méthode examine juste les parties pertinentes du corps.

- *Algorithmes basés sur les modèles d'apparences.* Contrairement aux algorithmes 3D, d'autres travaux tentent d'utiliser uniquement les apparences 2D de l'action. Une action est décrite par une séquence d'instances/poses 2D de l'objet. De nombreuses méthodes nécessitent une image normalisée de l'objet (généralement sans fond). Par exemple, Cui et al. [Cui 1995], Darrell et Pentland [Darrell 1993] et, aussi, Wilson et Bobick [Wilson 1995] présentent les résultats en utilisant des actions (principalement les gestes de la main), où les images en niveaux de gris (sans fond) sont utilisées. Bien que les apparences de la main restent assez semblables chez plusieurs personnes, à l'exception évidente de la couleur de la peau, les actions qui comprennent l'apparition de l'ensemble du corps ne sont pas aussi cohérentes visuellement chez différentes personnes en raison des variations naturelles et des apparences vestimentaires différentes.

- *Algorithmes basés sur les modèles de mouvements.* Ces approches tentent de caractériser le mouvement sans se référencier à des poses statiques du corps. Les auteurs de [Little 1995] utilisent le mouvement répétitif comme un signal d'avertissement fort pour reconnaître les mouvements de la marche cyclique. Ils permettent de suivre et de reconnaître les gens marchant dans des scènes extérieures en recueillant un vecteur caractérisant tout le corps. Ce vecteur emporte des

caractéristiques de mouvement de bas niveau et des mesures de périodicité. D'autres travaux, comme [Essa 1997], se concentrent sur les mouvements associés aux expressions faciales à l'aide des propriétés de mouvement se basant sur des régions prédéfinies. Le but de cette recherche est de reconnaître les expressions faciales humaines comme un système dynamique, où le mouvement des régions d'intérêt est pertinent. Ces approches caractérisent les expressions en utilisant les propriétés des mouvements sous-jacents plutôt que représenter l'action comme une séquence de poses.

Après avoir caractérisé le comportement, ses modèles sont analysés pour être reconnus. À présent, les comportements reconnus sont principalement : les mouvements des têtes et des membres et les gestes. Il y a deux types d'algorithmes de reconnaissance de comportements, comme suit :

- ***Méthode de Template Matching.*** L'idée de base est d'extraire des caractéristiques à partir des séquences vidéo, puis les comparer avec les modèles de comportement enregistrés à l'avance. Cette méthode a un faible cout de calcul, mais sensible au bruit.

- ***Méthode de l'espace d'états***. Chaque geste statique est défini comme un état, puis tous ces états sont combinés avec une probabilité. Chaque comportement est considéré comme un ensemble d'états. La classification du comportement dépend de la valeur maximale de la probabilité conjointe. Cette méthode exige un calcul itératif complexe.

4. Modélisation d'objets en mouvement dans les systèmes de vidéosurveillance

La modélisation d'objets joue un rôle crucial dans le suivi visuel car elle caractérise un objet d'intérêt. La sélection d'un modèle efficace joue un rôle essentiel dans le suivi. Seules les informations définies par le modèle sont utilisées pour l'estimation de la trajectoire. Par conséquent, un mauvais choix de modèle conduit inévitablement à un mauvais suivi. La gamme de modèles d'objets englobe divers types et dépend de l'application. Certaines applications requièrent juste un simple modèle, tandis que d'autres exigent un modèle précis et complexe pour réaliser le suivi [Jalal 2012].

Les représentations des formes d'objet, généralement utilisées pour le suivi sont : les points, les formes géométriques primitives, les silhouettes, les contours, les formes articulées et les squelettes [Yilmaz 2006], comme indiqué dans la Figure 6.

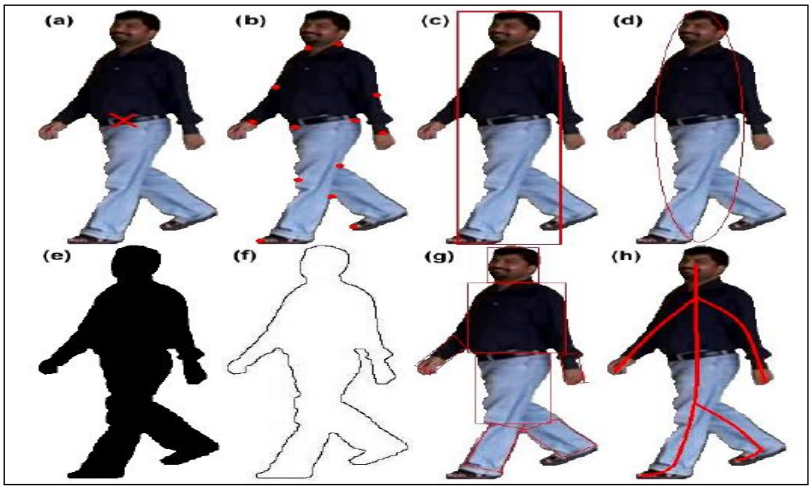

Figure 6. Les représentations de formes de l'objet (a) le point, (b) plusieurs points, (c) la forme géométrique primitive (rectangulaire), (d) la forme géométrique primitive (elliptique), (e) silhouette, (f) contour, (g) forme articulée, (h) squelette

– *Point.* Dans le suivi, la forme triviale est le point. Il s'agit d'une représentation simple de la localisation de l'objet. Un objet peut être représenté par un point qui peut être par exemple : son centre de masse, le centre de sa boite englobante, ou un point caractéristique de la forme. La représentation du point est adaptée pour les objets qui occupent des petites régions de l'image. Le suivi d'un point atténue également l'incertitude concernant la position de l'objet d'intérêt car il est fondé sur un seul point. Cette représentation peut s'étendre à un ensemble de points auxquels peuvent être associés des descripteurs locaux de couleur, de texture ou de mouvement. La représentation en point a été utilisée dans une multitude d'applications en raison de sa simplicité de traitement et la facilité de manipuler un point avec des algorithmes complexes [Veenman 2001]. Par exemple, elle a été utilisée pour le suivi dans l'imagerie radar [Novak 1981], le suivi distribué.

– *Formes géométriques.* La représentation en point d'un objet est un modèle simple. Cependant, il ne saisit pas toute la dynamique de l'objet. Pour exemple, la rotation n'est pas prise en charge avec la représentation en point. Par conséquent, des modèles plus avancés sont nécessaires pour répondre à ces types de problèmes. Les formes les plus populaires sont les formes géométriques primitives telles que les rectangles, les

carrés, les ellipses et les cercles. Le mouvement des objets associés à ces représentations est généralement modélisé à l'aide de transformation affine ou projective, translation, etc. Elles sont plus appropriées pour représenter des objets simples et rigides. Cependant, elles sont également utilisées pour des objets non-rigides. La représentation en rectangle est omniprésente dans le suivi géométrique d'objets tels que les voitures [Melo 2006, Ching 2003] ou le suivi d'objets à faible distorsion tels que les personnes [Yang 2005]. Une forme carrée adaptative a été utilisée pour la représentation de l'objet dans [Bradski 1998]. L'ellipse offre l'avantage d'arrondir les bords par rapport au rectangle lorsque l'objet ne possède pas d'arêtes vives [Chang 2005]. Dans [Comaniciu 2000, Comaniciu 2003], l'auteur a utilisé une forme elliptique pour représenter l'objet en mouvement.

- *Modèles articulés.* Les formes articulées sont utilisées pour le suivi si différentes parties de l'objet d'intérêt doivent être décrites individuellement (par exemple, les jambes, les bras et la tête). Ce type de représentation est beaucoup adapté pour le corps humain, qui est un objet articulé avec la tête, les mains, les jambes, etc. Ces éléments constitutifs devraient être liés par un modèle cinématique. Les parties constitutives peuvent être représentées par des formes géométriques primitives telles que des rectangles, des cercles et des ellipses. Ramanan et al. ont développé un modèle de forme articulé pour décrire la configuration du corps [Ramanan 2003]. Dans [Haritaoglu 2000], les positions des différents membres du corps sont utilisées pour analyser le comportement des gens.

- *Squelettes.* Dans cette représentation, le squelette d'un objet peut être extrait pour modéliser les objets articulés et rigides. On peut définir le squelette comme un ensemble d'articulations qui décrit les dépendances et définit les contraintes entre les représentations des parties. Dans [Ali 2001], l'auteur a utilisé le modèle squelettique pour la segmentation automatique et la reconnaissance de l'activité humaine continue.

- *Silhouettes.* La zone à l'intérieur du contour représente la silhouette d'un objet. C'est un masque binaire dense qui représente un objet d'intérêt. Les représentations en silhouette sont adaptées pour le suivi des formes complexes non rigides.

- *Contours.* Dans cette représentation, la limite d'un objet est définie comme un contour. Au lieu de stocker l'ensemble de la silhouette, les contours décrivent seulement les bords entourant l'objet. Une forme non rigide de l'objet peut être mieux décrite par ces représentations [Yilmaz 2004].

5. Scalabilité dans les systèmes de vidéosurveillance

La scalabilité est la capacité d'un système de pouvoir traiter un volume croissant de tâches ou la capacité de se développer pour s'adapter au rythme de la demande [Bondi 2000]. La scalabilité est généralement difficile à définir sans préciser le contexte du système [Hill 1990]. Dans le domaine de traitement de signal, la scalabilité est la capacité de représenter un signal à différents niveaux d'information. Le besoin d'un codage vidéo scalable provient de l'évolution continue des dispositifs de réception et de l'utilisation croissante des systèmes ayant des qualités de transmission variables. Le codage vidéo d'aujourd'hui est utilisé dans une large gamme d'applications telles que : la messagerie multimédia, la téléphonie vidéo et la visioconférence, le streaming vidéo, etc. En particulier, l'internet et les réseaux sans fil gagnent de plus en plus d'importance pour les applications vidéo. Cependant, la transmission vidéo dans de tels systèmes est exposée à des conditions de transmission variables. Ceci peut être remédié à l'aide des avantages de la scalabilité. En outre, les contenus vidéo sont livrés à une variété de dispositifs de décodage avec des capacités d'affichage et de calcul différentes. Dans ces environnements hétérogènes, une adaptation flexible d'un contenu codé une seule fois est souhaitable, aussi une interopérabilité du codeur et du décodeur produits par différents fabricants. Les questions de la scalabilité ont d'abord été abordées dans les normes MPEG-2, H.263 et MPEG-4, puis complètement répondues dans la norme H.264/SVC, comme une extension de la norme H.264/AVC. Un flux vidéo est appelé scalable lorsque des parties peuvent être éliminées de façon que le sous-flux résultant forme un autre flux valide pour certains décodeurs, et représente le contenu de la source avec une qualité de reconstruction qui est moins que celle initiale complète mais elle est élevée si on considère la qualité inférieure des données restantes. Le flux d'un codage vidéo scalable SVC (Scalable Video Coding) se compose d'une couche de base et des couches de rehaussement. Comme le montre la Figure 7, chaque couche de rehaussement améliore la vidéo selon l'une des trois échelles : temporelle, spatiale et qualité.

- *La scalabilité temporelle.* La scalabilité temporelle permet d'ajuster la fréquence temporelle d'un flux vidéo afin de réduire l'intervalle entre les deux trames

consécutives et améliorer la fluidité du mouvement. Elle est fournie par la structure hiérarchique du GOP (Group Of Pictures).

- *La scalabilité spatiale.* La scalabilité spatiale est obtenue en utilisant une structure de résolution pyramidale. Un flux scalable contient plusieurs couches de différentes résolutions spatiales. La résolution de la couche inférieure est appelée couche de base. La couche la plus élevée est appelée couche d'amélioration. Elle peut alors utiliser les informations de la couche de base à l'aide des mécanismes de prédiction inter-couches qui exploitent les redondances entre les différents niveaux du flux.

- *La scalabilité SNR (qualité).* Avec la scalabilité de la qualité, le sous-flux offre la même résolution spatio-temporelle que le flux complet, mais avec une qualité inférieure. La scalabilité de la qualité est aussi communément appelée la scalabilité de fidélité ou SNR (rapport signal sur bruit).

Deux autres modes de scalabilité sont rarement utilisées : la scalabilité de régions d'intérêt et la scalabilité orientée objet, dans lesquels les sous-flux représentent généralement des régions spatialement contigus de l'image originale. Ces différents modes de scalabilité peuvent être également combinés, de sorte qu'un flux scalable unique peut intégrer le contenu de la source avec différents débits et résolutions spatio-temporelles [Schwarz 2007].

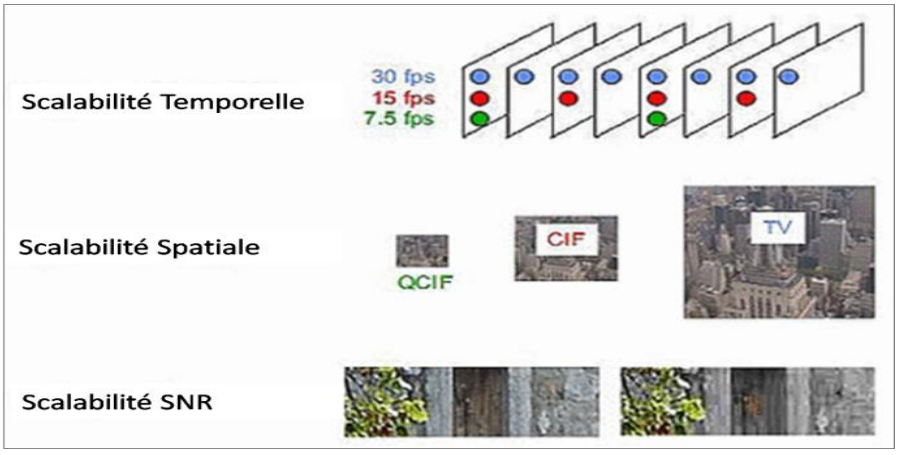

Figure 7. Les échelles de base de la scalabilité dans le codage vidéo

6. Motivations

Les systèmes de vidéosurveillance se sont largement développés ces dernières années, surtout après l'apparition de la vidéosurveillance intelligente. L'objectif de cette dernière est: obtenir une description de ce qui se passe dans une zone surveillée, puis prendre l'action appropriée fondée sur cette interprétation. Comme nous l'avons remarqué dans la Section 2.4, malgré les différents types de systèmes de vidéosurveillance existants, ils suivent tous la même architecture décrite dans la Section 2.3. De ce fait, le processus de traitement des flux vidéo capturés se déroule de façon similaire dans tous les systèmes quelques soit leurs contextes applicatifs ou leurs applications visées. Conscients de l'importance d'avoir une chaine de traitement de flux vidéo spécifique pour chaque système de surveillance selon son application exigée et devant la diversité des fonctions analytiques (citées dans la Section 3), nous nous sommes intéressés dans cette thèse à l'aspect applicatif des systèmes de vidéosurveillance. Ainsi, comme idée introductive, nous avons commencé par suggérer, dans la Section 2.4.4, une nouvelle classification des systèmes selon leurs niveaux applicatifs estimés.

Comme décrit tout au long de ce chapitre, les travaux de la littérature se sont massivement concentrés sur l'étude et le développement de méthodologies pour assurer et améliorer les tâches analytiques des systèmes de vidéosurveillance. Néanmoins, les performances des systèmes de surveillance peuvent être améliorées non seulement en agissant sur la partie analytique mais aussi sur les autres phases. En fait, dans les travaux connexes, le flux vidéo enregistré est intégralement traité alors que réellement juste une faible partie de la scène est utile. Dans ce contexte, nous proposons, dans le Chapitre 2, d'intégrer une étape de pré-analyse dans l'architecture des systèmes de surveillance qui sélectionne et extrait les régions intéressantes de la séquence capturée suivant l'application de surveillance demandée, ce qui permettra d'analyser la scène à moindre coûts computationnels .

Une autre limitation, qui peut être distinguée depuis les systèmes existants de la vidéosurveillance, est le fait qu'ils sont à un seul niveau applicatif. Ainsi, chaque système développé jusqu'à présent réalise une seule application précise. Notre idée est de proposer une nouvelle architecture incrémentale qui permet de concevoir des systèmes répondant à plusieurs applications à la fois. Cette architecture exploite la phase de pré-analyse suggérée et la multitude des fonctions analytiques existantes. Elle intègre aussi le concept de la scalabilité dans sa démarche. L'utilisation de la scalabilité, dans les travaux connexes de surveillance,

était toujours à travers la phase d'encodage. Dans la contribution décrite dans le Chapitre 3, nous adaptons le concept de la scalabilité pour des fins applicatifs.

7. Conclusion

Dans ce chapitre, nous avons montré en présentant la situation mondiale et donnant quelques chiffres explicatifs que les systèmes de vidéosurveillance sont en plein développement et, représente un domaine de recherche de plus en plus actif. Nous avons détaillé les systèmes de vidéosurveillance en décrivant leurs architectures et citant, de manière non exhaustive, leurs classifications possibles. Nous avons proposé une nouvelle classification basée sur le niveau d'application des systèmes de surveillance. Nous avons, aussi, fait un tour d'horizon sur les méthodes couramment utilisées pour les différentes étapes analytiques d'un système de vidéosurveillance. Les rôles de la modélisation d'objets en mouvement et de la scalabilité dans la vidéosurveillance ont été, également, discutés. A l'issue de cet état de l'art, nous avons donné les limitations des travaux de surveillance existants pour justifier les motivations du travail de thèse.

Chapitre 2

Une nouvelle approche de pré-analyse vidéo pour les systèmes de vidéosurveillance

1. Introduction

Ce chapitre décrit l'approche proposée pour la pré-analyse dans les systèmes de vidéosurveillance. Il s'agit de simplifier les sections susceptibles de contenir des informations d'intérêt. Pour être efficace, il faut réduire la quantité de données, mais ne pas détruire les informations utiles pour une analyse ultérieure. Ce processus de simplification est précédé par une étape d'extraction de régions d'intérêt pour détecter les parties requises. La simplification est réalisée en deux méthodes distinctes.

La première section de ce chapitre (Section 2) fournit une description de l'approche générale proposée de pré-analyse. La Section 3 présente l'algorithme d'extraction de régions d'intérêt utilisé par notre approche. Les Sections 4 et 5 décrivent les deux méthodes de pré-analyse pour les objets en mouvement. Quant à la dernière section (Section 6) de ce chapitre, elle évalue les performances et discute les résultats obtenus.

2. Approche générale proposée

Dans le cadre des applications de vidéosurveillance, l'objectif consiste, le plus souvent, à être capable de détecter les événements d'intérêts tels que, par exemple, la détection d'accident, de vols, ou la présence de personnes non autorisées. L'utilisation classique de ces scènes enregistrées consiste soit à avoir un opérateur humain surveillant les écrans directement reliés au capteur via un réseau haut débit, soit à utiliser a posteriori les vidéos pour, par exemple, retrouver l'auteur d'un vol. Pour permettre une utilisation plus performante des flux vidéos disponibles, il est nécessaire de concevoir des méthodes capables de détecter automatiquement les évènements d'intérêts et de disposer d'une méthode de compression efficace (type H.264 par exemple) lorsque le débit disponible ne permet pas de transmettre les vidéos brutes en temps réel. Il s'agit donc d'enregistrer, compresser, transmettre, décompresser et analyser la vidéo en temps réel. Néanmoins, dans de nombreux cas, seule une très faible proportion des informations transmises est réellement utile, c'est-à-dire les zones dans certaines portions de la vidéo montrant un évènement d'intérêt. Pour réduire le flux de données vidéo avant d'être transmis, une solution consiste à analyser la scène vidéo au niveau du capteur, et à ne transmettre après une éventuelle compression que les informations utiles (les régions d'intérêt). Les algorithmes d'analyse pouvant être complexes, une telle solution peut nécessiter des moyens de calcul importants au niveau du capteur ce qui n'est pas souvent possible. Pour remédier à ce problème, notre solution

consiste à détecter, grâce à une phase de pré-analyse, au niveau de la caméra les zones susceptibles de contenir des informations utiles, puis à analyser ces zones pour éliminer l'information a priori inutile. Finalement, seules ces zones simplifiées seront compressées pour être analysées plus finement au niveau de la réception. Pour être efficace, la phase de pré-analyse au niveau du capteur doit être relativement légère en termes de charge de calcul, et ne pas détruire l'information utile pour la détection d'évènement. Par conséquent, le schéma habituel des systèmes de vidéosurveillance est amélioré, comme le montre la Figure 8, pour devenir : enregistrer, pré-analyser, compresser, transmettre, décompresser et analyser la scène. Ajouter la phase de pré-analyse au processus habituel demande des calculs supplémentaires au niveau du capteur. De retour, il accélère la transmission et l'analyse des données. Dans l'architecture matérielle typique, les flux vidéos sont stockés sur un serveur central et ensuite distribués pour l'analyse. Puisque dans notre approche proposée la pré-analyse doit être réalisée du côté du capteur, cette architecture n'est plus adéquate. Comme présenté par la Figure 9, une architecture distribuée où le stockage et la pré-analyse de vidéo sont effectués au niveau du capteur est adaptée [Senior 2009]. Les informations sont transmises uniquement lorsqu'elles sont demandées par l'utilisateur.

Lors de la pré-analyse, le processus de simplification est précédé par une étape d'extraction de régions d'intérêt pour détecter les zones décrivant les mouvements d'intérêt, comme le montre la Figure 10. Par la suite, ces régions sont simplifiées pour ne garder que les informations intéressantes. La pré-analyse vidéo est réalisée au moyen de deux méthodes différentes. Une première de filtrage qui simplifie les données d'une manière spatio-temporelle pour ne garder que les détails nécessaires pour une analyse ultérieure plus fine. Une deuxième de modélisation géométrique des objets en mouvement pour les représenter en formes parallélépipédiques avec le maintien des caractéristiques requises pour le suivi d'objets. Ces deux techniques font parties de l'architecture scalable orientée vers l'application proposée, dans le Chapitre 3, pour les systèmes de vidéosurveillance.

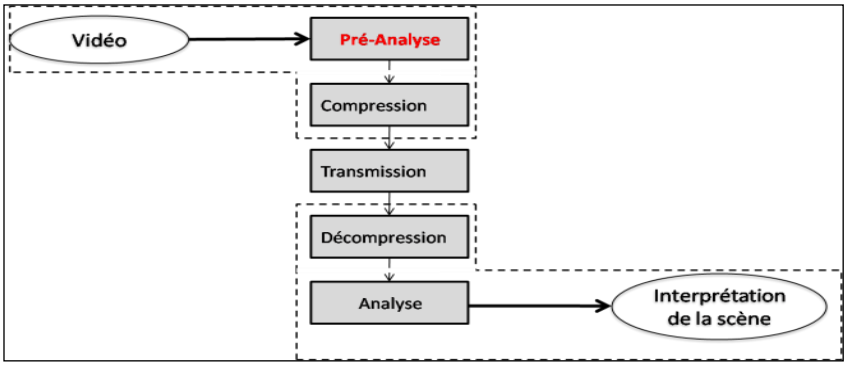

Figure 8. Architecture améliorée proposée des systèmes de vidéosurveillance

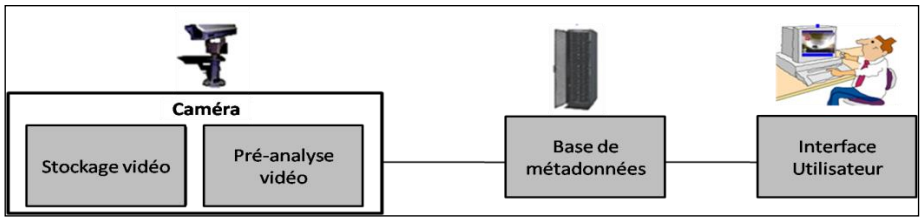

Figure 9. Dans l'architecture proposée, la pré-analyse et le stockage vidéo sont réalisés au niveau de la caméra

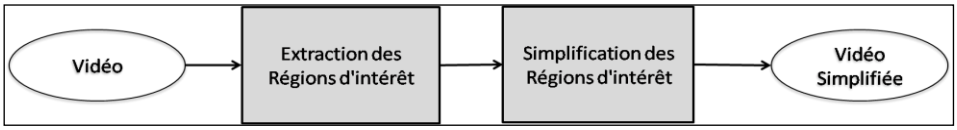

Figure 10. Les deux étapes principales de la pré-analyse vidéo

3. Extraction des régions d'intérêt

Généralement, dans les applications de vidéosurveillance, la séparation entre le fond et les objets en mouvement est une étape cruciale. La soustraction de fond est l'une des techniques clés pour l'analyse automatique de vidéo, en particulier dans les applications de la vidéosurveillance reposant sur une caméra fixe. Une segmentation robuste et précise simplifie les étapes de traitement suivantes. L'extraction des régions d'intérêt est essentielle pour la pré-analyse. Les régions nécessaires sont les zones pleines de renseignements sur les événements d'intérêt. Puisque le contexte de notre travail est la surveillance routière et des lieux publics, les zones contenant des objets mobiles tels que des piétons, des véhicules représentent les régions pertinentes. Dans la Figure 11, la phase d'extraction des régions d'intérêt combine deux grandes étapes : la soustraction de fond et l'extraction des blobs.

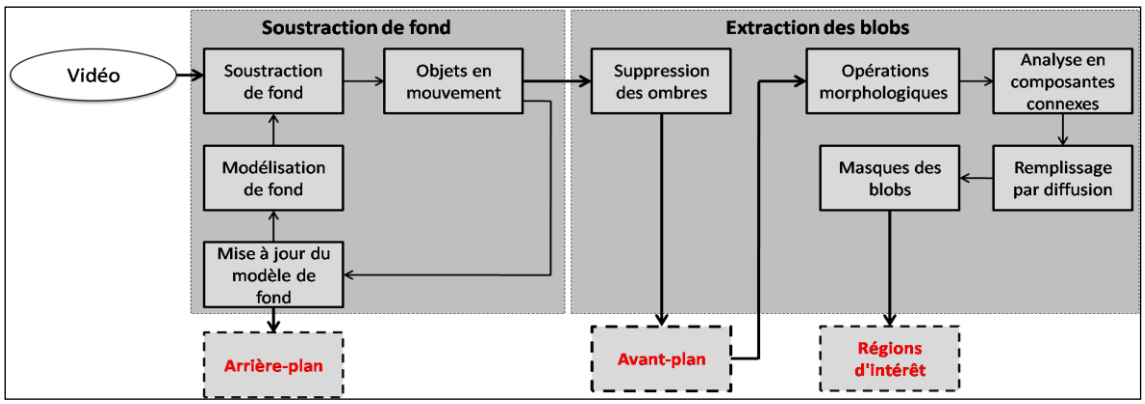

Figure 11. Les deux étapes d'extraction des régions d'intérêt : Soustraction de fond et Extraction des blobs

3.1. *Soustraction de fond*

Obtenir une soustraction de fond précise et peu coûteuse en temps de calcul dans le contexte de la vidéosurveillance est un problème réel. Les scènes utilisées sont des scènes de vidéosurveillance extérieures comportant plusieurs perturbations telles que les changements de luminosité (forts éclairages, faibles éclairages, ombres), la présence d'arrière-plan dynamique (mouvements de branches d'arbres), les occlusions,.... Elles sont aussi des scènes enregistrées avec une seule caméra supposée fixe avec un fond statique.

Nous avons fait recours à une technique récursive de soustraction d'arrière-plan [Zivkovic 2004] qui est une version améliorée de la méthode connue des mélanges de gaussiennes (GMM) de Stauffer et al. [Stauffer 1999]. Des équations récursives sont intégrées pour sélectionner constamment le nombre approprié de composantes pour chaque pixel. Ce choix, étant une procédure en temps réel, permet à l'algorithme de s'adapter automatiquement et pleinement à la scène. Les résultats de segmentation et le temps de traitement sont améliorés par rapport à la méthode d'origine, comme expliqué dans l'étude comparative menée dans le travail de [Brutzer 2011]. En outre, cette technique est robuste face aux changements d'éclairage ralentis, aux mouvements périodiques avec un fond encombré, aux mouvements lents des objets, aux changements de la scène à long terme, et aux bruits de la caméra [Zivkovic 2006].

Construction du modèle de fond

La valeur d'un pixel à l'instant $\boldsymbol{t}$ dans n'importe quel espace de couleurs est notée $\vec{\boldsymbol{x}}^{(\boldsymbol{t})}$. La soustraction de fond à base de pixel décide pour chaque pixel s'il appartient au fond (BG) ou à un objet de l'avant-plan (FG). La décision Bayésienne $\boldsymbol{R}$ est calculée par :

$$\boldsymbol{R} = \frac{\boldsymbol{p}(\boldsymbol{BG}|\vec{\boldsymbol{x}}^{(\boldsymbol{t})})}{\boldsymbol{p}(\boldsymbol{FG}|\vec{\boldsymbol{x}}^{(\boldsymbol{t})})} = \frac{\boldsymbol{p}(\vec{\boldsymbol{x}}^{(\boldsymbol{t})}|\boldsymbol{BG})\boldsymbol{p}(\boldsymbol{BG})}{\boldsymbol{p}(\vec{\boldsymbol{x}}^{(\boldsymbol{t})}|\boldsymbol{FG})\boldsymbol{p}(\boldsymbol{FG})} \qquad (20)$$

Généralement, aucune information sur la dynamique des objets en mouvement dans la scène ne peut être connue au préalable. Par conséquent, on suppose que $\boldsymbol{p}(\boldsymbol{FG}) = \boldsymbol{p}(\boldsymbol{BG})$ et que $\boldsymbol{p}(\vec{\boldsymbol{x}}^{(\boldsymbol{t})}|\boldsymbol{FG}) = \boldsymbol{c}_{\boldsymbol{FG}}$ est une distribution uniforme pour l'apparition des objets d'avant-plan. Un pixel appartient au fond si :

$$\boldsymbol{p}(\vec{\boldsymbol{x}}^{(\boldsymbol{t})}|\boldsymbol{BG}) > \boldsymbol{c}_{\boldsymbol{thr}}\left(= \boldsymbol{p}(\vec{\boldsymbol{x}}^{(\boldsymbol{t})}|\boldsymbol{FG})\boldsymbol{p}(\boldsymbol{FG})/\boldsymbol{p}(\boldsymbol{BG})\right) \qquad (21)$$

avec c_{thr} une valeur seuil et $p(\vec{x}^{(t)}|BG)$ est le modèle de fond. Ce modèle est estimé depuis un ensemble d'apprentissage nommé χ. Le modèle estimé est désigné par $p(\vec{x}^{(t)}|\chi, BG)$ et dépend de l'ensemble d'apprentissage.

Estimation du modèle de fond

Dans la pratique, l'éclairage de la scène peut changer progressivement ou soudainement. Un nouvel objet peut s'introduire dans la scène ou un objet déjà présent peut se retirer. Afin de s'adapter à ces changements, l'ensemble d'apprentissage est mis à jour en ajoutant des nouveaux échantillons et écartant les anciens. Une période de temps raisonnable T est choisie avec à l'instant t, $\chi_t = \{x^{(t)}, \dots, x^{(t-T)}\}$. Pour chaque nouveau échantillon, l'ensemble d'apprentissage χ_t est mis à jour et le modèle $p(\vec{x}^{(t)}|\chi_t, BG)$ est re-estimé. Cependant, parmi les échantillons récents, certaines valeurs peuvent appartenir à des objets de l'avant-plan et cette estimation est désignée par $p(\vec{x}^{(t)}|\chi_t, BG + FG)$. Pour une GMM avec M composantes :

$$p(\vec{x}^{(t)}|\chi_t, BG + FG) = \sum_{m=1}^{M} \hat{\pi}_m N(\vec{x}; \hat{\vec{\mu}}_m, \hat{\sigma}_m^2 I) \quad (22)$$

avec $\hat{\vec{\mu}}_1, \dots, \hat{\vec{\mu}}_M$ les estimations des moyennes et $\hat{\sigma}_1, \dots, \hat{\sigma}_M$ les estimations des variances qui décrivent les composantes Gaussiennes. La matrice de covariance est diagonale et I est la matrice d'identité. Les valeurs des poids des mélanges $\hat{\pi}_m$ sont non négatives. Étant donné un nouvel échantillon $\vec{x}^{(t)}$ à l'instant t, les équations récursives de mise à jour sont :

$$\hat{\pi}_m \leftarrow \hat{\pi}_m + \alpha\left(o_m^{(t)} - \hat{\pi}_m\right) \quad (23)$$

$$\hat{\vec{\mu}}_m \leftarrow \hat{\vec{\mu}}_m + o_m^{(t)}(\alpha/\hat{\pi}_m)\vec{\delta}_m \quad (24)$$

$$\hat{\sigma}_m^2 \leftarrow \hat{\sigma}_m^2 + o_m^{(t)}(\alpha/\hat{\pi}_m)\left(\vec{\delta}_m^T\vec{\delta}_m - \hat{\sigma}_m^2\right) \quad (25)$$

avec $\vec{\delta}_m = \vec{x}^{(t)} - \hat{\vec{\mu}}_m$. Au lieu d'utiliser la période de temps T, le paramètre α décrit une enveloppe à décroissance exponentielle utilisée pour limiter l'influence des échantillons anciens et donc, $\alpha \simeq 1/T$. Pour chaque nouvel échantillon, la valeur de la propriété $o_m^{(t)}$ est égale à 1 pour la composante la plus proche et s'annule pour les autres. La distance calculée entre l'échantillon et la composante est la distance de Mahalanobis. La distance carrée par rapport à la m-ème composante est calculée par : $D_m^2(\vec{x}^{(t)}) = \vec{\delta}_m^T\vec{\delta}_m/\hat{\sigma}_m^2$. Si aucune proche composante n'existe, une nouvelle est générée avec $\hat{\pi}_{M+1} = \alpha$, $\hat{\vec{\mu}}_{M+1} = \vec{x}^{(t)}$ et $\hat{\sigma}_{M+1} = \sigma_0$ où σ_0 est une variance initiale appropriée. Si le nombre maximal de composantes est atteint, celle avec la plus faible valeur de $\hat{\pi}_M$ est écartée.

L'algorithme décrit présente un algorithme de classification en temps réel. Habituellement, les objets intrus de premier plan sont représentés par des composantes supplémentaires avec de faibles valeurs de poids $\hat{\pi}_M$. Par conséquent, le modèle de fond peut être approximé par les B premières composantes :

$$p(\vec{x}^{(t)}|\chi_t, BG) \sim \sum_{m=1}^{B} \hat{\pi}_m N(\vec{x}; \hat{\vec{\mu}}_m, \hat{\sigma}_m^2 I) \qquad (26)$$

Les composantes sont triées selon des poids décroissants $\hat{\pi}_M$:

$$B = argmin_b \left(\sum_{m=1}^{b} \hat{\pi}_m > (1 - c_f)\right) \qquad (27)$$

Avec $\mathbf{c_f}$ est une mesure de la partie maximale des données qui peuvent appartenir à des objets de premier plan sans influencer sur le modèle d'arrière-plan. Par exemple, si un nouvel objet entre dans une scène et reste statique pendant un certain temps, une composante supplémentaire est générée avec un poids π_{B+1} qui augmente constamment. Si l'objet reste statique assez longtemps, son poids devient supérieur à c_f et il peut être considéré comme faisant partie de l'arrière-plan. D'après l'équation (23), l'objet doit être statique pendant environ $\log(1 - c_f) / \log(1 - \alpha)$ images. Par exemple pour $c_f = 0.1$ et $\alpha = 0.001$, le nombre d'images est 105.

Sélection du nombre de composantes du modèle de fond

Le poids π_M décrit la quantité de données appartenant à la m-ème composante du modèle des mélanges de Gaussiennes. Il peut être considéré comme la probabilité que l'échantillon provient de la m-ème composante et de cette façon le $\pi_M - s$ définit une distribution multinomiale sous-jacente. Supposons avoir t échantillons et chacun d'entre eux appartient à l'une des composantes de la GMM. Supposons également que le nombre d'échantillons appartenant à la m-ème composante est : $n_m = \sum_{i=1}^{t} o_m^{(i)}$. La distribution multinomiale supposée pour $n_m - s$ possède une fonction de vraisemblance $\mathcal{L} = \prod_{m=1}^{M} \pi_m^{n_m}$. La somme des poids est égale à 1. Un multiplicateur de Lagrange λ est introduit et le maximum de vraisemblance est estimé : $\frac{\partial}{\partial_{\hat{\pi}_m}} \left(\log \mathcal{L} + \lambda(\sum_{m=1}^{M} \hat{\pi}_m - 1)\right) = 0$ (28)

Après une simplification, l'équation (28) devient : $\hat{\pi}_m^{(t)} = \frac{n_m}{t} = \frac{1}{t} \sum_{i=1}^{t} o_m^{(i)}$ (29)

L'estimation à partir des t échantillons est désignée comme $\hat{\pi}_m^{(t)}$. Elle peut être reformulée dans une forme récursive comme suit :

$$\hat{\pi}_m^{(t)} = \hat{\pi}_m^{(t-1)} + 1/t(o_m^{(t)} - \hat{\pi}_m^{(t-1)}) \qquad (30)$$

Si la valeur de $^{\mathbf{1}}/_{\boldsymbol{t}}$ est égale à $\boldsymbol{\alpha = 1/T}$, l'influence des nouveaux échantillons sera fixée et l'équation (23) est réobtenue. Fixer l'influence des nouveaux échantillons signifie qu'ils ont plus d'importance et que la contribution des anciens échantillons est pondérée d'une manière exponentielle décroissante, comme mentionné précédemment. Une connaissance préalable de la distribution multinomiale peut être introduite en utilisant une densité a priori de Dirichlet $\mathrm{P} = \prod_{\mathbf{m=1}}^{\mathbf{M}} \boldsymbol{\pi}_{\mathbf{m}}^{\mathbf{c_m}}$. Les coefficients $\mathbf{c_m} = -\mathbf{c}$ sont des coefficients négatifs ce qui signifie que la composante $\mathbf{m}$ existe seulement s'il y a suffisamment de preuves à partir des données de l'existence de cette classe. La solution de l'estimateur du maximum a posteriori (MAP) résulte de $\frac{\partial}{\partial_{\hat{\pi}_\mathbf{m}}}\left(\mathbf{\log \mathcal{L} + \log P} + \boldsymbol{\lambda}(\sum_{\mathbf{m=1}}^{\mathbf{M}} \hat{\boldsymbol{\pi}}_\mathbf{m} - \mathbf{1})\right) = \mathbf{0}$, où $\mathrm{P} = \sum_{\mathbf{m=1}}^{\mathbf{M}} \boldsymbol{\pi}_{\mathbf{m}}^{-\mathbf{c}}$.

Alors, l'équation suivante est obtenue :

$$\hat{\boldsymbol{\pi}}_{\boldsymbol{m}}^{(\boldsymbol{t})} = \frac{\mathbf{1}}{\boldsymbol{K}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{o}_{\boldsymbol{m}}^{(\boldsymbol{i})} - \boldsymbol{c}\right) \qquad (31)$$

avec $\boldsymbol{K} = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{o}_{\boldsymbol{m}}^{(\boldsymbol{t})} - \boldsymbol{c}\right) = \boldsymbol{t} - \boldsymbol{Mc}$. L'équation (31) peut être reformulée comme suit :

$$\hat{\boldsymbol{\pi}}_{\boldsymbol{m}}^{(\boldsymbol{t})} = \frac{\hat{\boldsymbol{\Pi}}_{\boldsymbol{m}} - \boldsymbol{c/t}}{\mathbf{1} - \boldsymbol{Mc/t}} \qquad (32)$$

où $\hat{\boldsymbol{\Pi}}_{\boldsymbol{m}} = \frac{\mathbf{1}}{\boldsymbol{t}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{o}_{\boldsymbol{m}}^{(\boldsymbol{t})}$ est l'estimation du maximum de vraisemblance de l'équation (29). La version récursive de (31) avec une valeur fixe de $\boldsymbol{c_T} = \boldsymbol{c/t}$ est :

$$\hat{\boldsymbol{\pi}}_{\boldsymbol{m}}^{(\boldsymbol{t})} = \hat{\boldsymbol{\pi}}_{\boldsymbol{m}}^{(\boldsymbol{t-1})} + \mathbf{1/t}\left(\frac{\boldsymbol{o}_{\boldsymbol{m}}^{(\boldsymbol{t})}}{\mathbf{1} - \boldsymbol{Mc_T}} - \hat{\boldsymbol{\pi}}_{\boldsymbol{m}}^{(\boldsymbol{t-1})}\right) - \mathbf{1/t}\frac{\boldsymbol{c_T}}{\mathbf{1} - \boldsymbol{Mc_T}} \qquad (33)$$

Puisque un nombre faible de composantes est attendu et la valeur de $\boldsymbol{c_T}$ est faible, la valeur de $\mathbf{1} - \boldsymbol{Mc_T}$ peut être approximée à 1. En attribuant $\frac{\mathbf{1}}{\boldsymbol{t}}$ à la valeur de $\boldsymbol{\alpha}$, l'équation adaptative finale de mise à jour est obtenue :

$$\hat{\boldsymbol{\pi}}_{\boldsymbol{m}} \leftarrow \hat{\boldsymbol{\pi}}_{\boldsymbol{m}} + \boldsymbol{\alpha}\left(\boldsymbol{o}_{\boldsymbol{m}}^{(\boldsymbol{t})} - \hat{\boldsymbol{\pi}}_{\boldsymbol{m}}\right) - \boldsymbol{\alpha c_T} \qquad (34)$$

Cette équation remplace l'équation (23). Après chaque mise à jour, les valeurs de $\boldsymbol{\pi_m - s}$ sont normalisées afin que leur somme soit égale à $\mathbf{1}$. La GMM débute avec une seule composante centrée autour du premier échantillon, ensuite des nouvelles composantes s'ajoutent, comme mentionné précédemment. La densité à priori de Dirichlet avec des coefficients négatifs supprime les composantes qui ne sont pas supportées par des échantillons. La composante $\boldsymbol{m}$ est écartée lorsque son poids $\boldsymbol{\pi_m}$ devient négatif ce qui permet de garantir que les poids du GMM restent non négatifs. La Figure 12 montre quelques exemples de modélisation de plusieurs scènes avec un nombre variable de composantes. Pour une valeur choisie de $\boldsymbol{\alpha = 1/T}$, la valeur minimale d'échantillons exigée pour supporter une composante est $\boldsymbol{c = 0.01 * T}$. Par conséquent, la valeur $\boldsymbol{c_T = 0.01}$.

Figure 12. Exemples de modélisation de fond dans différentes scènes avec la version améliorée de GMM [Zivkovic 2004]. Le modèle de mélange comporte un nombre variable de composantes gaussiennes en fonction de la dynamique de la scène

3.2. *Extraction des blobs*

Chaque nouvelle image est comparée au modèle de l'arrière-plan construit afin d'en extraire les pixels couleurs s'écartant de la modélisation. Comme présenté dans la Figure 11, l'extraction de blobs permet d'assurer une segmentation plus précise des régions d'intérêt de la scène. En premier lieu, une étape de suppression des ombres est effectuée sur les zones en mouvement détectées lors de la soustraction de fond pour extraire les objets de l'avant-plan. Ensuite, les régions d'intérêt pertinentes sont calculées par l'application des opérateurs morphologiques, l'analyse par composantes connexes et le remplissage par diffusion.

3.2.1. Suppression des ombres

La présence des ombres est un problème qui se pose lors de la segmentation d'objets. Les ombres sont dues à l'interposition d'une source lumineuse, un objet de la scène et une surface sur laquelle se réfléchit cette lumière (voir Figure 13). Les ombres peuvent être classées en deux catégories : la partie non allumée de l'objet est appelée ombre propre (self-shadow); la zone projetée sur la scène par l'objet est appelée ombre portée (cast-shadow).

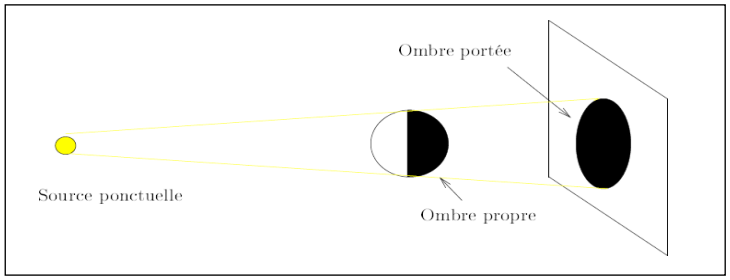

Figure 13. Les types d'ombre dans le cas d'un objet éclairé par une source de lumière ponctuelle

Dans notre contexte, nous nous intéressons seulement aux ombres portées car elles suivent les mouvements de l'objet (correctement appelée ombre portée en mouvement) et représentent des zones en mouvement à éliminer. Les ombres propres quant à elles, ne sont pas d'intérêt car elles se trouvent sur l'objet en mouvement et prennent sa forme dans sa partie non éclairée. Pour réussir la segmentation des régions d'intérêt, les ombres portées ne doivent pas être classées comme appartenant aux objets d'avant-plan. Malheureusement, les points des objets et leurs ombres associées partagent deux caractéristiques visuelles importantes : le mouvement et la détectabilité. L'absence de séparation entre les objets en mouvement et les ombres portées provoque des erreurs possibles. Ainsi, quelle que soit la mise à jour du fond, souvent les parties mobiles des objets et des ombres sont détectées en même temps et regroupées ensemble. Par conséquent, la forme de l'objet est altérée et les propriétés géométriques sont modifiées. Ce problème affecte de nombreuses tâches ultérieures, comme la classification d'objets. De plus, la probabilité de fausse segmentation d'objets (par exemple des objets fusionnés en un seul blob) augmente en raison de la connectivité via les ombres entre les différents objets ce qui provoque des erreurs dans les étapes ultérieures de suivi et d'identification d'objets.

Afin de remédier à ces inconvénients, nous avons utilisé une approche pour la détection et la suppression d'ombres basée sur les travaux de [Cucchiara 2001] qui consiste en une analyse des couleurs des pixels par rapport à celle de l'arrière-plan dans l'espace couleur HSV (Hue Saturation Value). L'espace HSV correspond étroitement à la perception humaine de la couleur et il a prouvé une meilleure précision dans la distinction de l'ombre par rapport à l'espace RGB (Red Green Blue). Seuls les points appartenant à des objets mobiles sont analysés. La méthode estime l'effet de l'occlusion en raison de l'ombre sur le changement des valeurs des composantes H, S et V. En fait, un point d'ombre assombrit le point d'arrière-plan sur lequel il est projeté, tandis qu'un point d'objet pourrait obscurcir ou non, en fonction de la texture et de la couleur de l'objet.

Le changement de l'apparence locale à cause de l'ombre est exploitée par le calcul du rapport d'apparence d'un pixel entre l'image actuelle et l'image de référence :

$$\boldsymbol{R_k}(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{s_{k+1}}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{s_k}(\boldsymbol{x},\boldsymbol{y})} \qquad (35)$$

où $\boldsymbol{s_k}$ est la luminance d'un point de l'image de coordonnées $(\boldsymbol{x},\boldsymbol{y})$ à l'instant $\boldsymbol{k}$.

Dans l'équation (35), la luminance $\boldsymbol{s_k}(\boldsymbol{x},\boldsymbol{y})$ est approximée par $\boldsymbol{I_k^V}(\boldsymbol{x},\boldsymbol{y})$ où $\boldsymbol{I_k^V}(\boldsymbol{x},\boldsymbol{y})$ est la valeur de l'intensité de la composante V de l'espace HSV du pixel de coordonnées $(\boldsymbol{x},\boldsymbol{y})$

dans l'image $\boldsymbol{k}$. Ainsi, un masque d'ombre $\boldsymbol{SP_k}$ est défini pour chaque pixel (appartenant à un objet en mouvement) avec trois conditions comme suit :

$$\boldsymbol{SP_k}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \mathbf{1}\ \boldsymbol{si} & \boldsymbol{\alpha} \le \frac{\boldsymbol{I_k^V(x,y)}}{\boldsymbol{B_k^V(x,y)}} \le \boldsymbol{\beta} \\ & \wedge \left(\boldsymbol{I_k^S(x,y)} - \boldsymbol{B_k^S(x,y)}\right) \le \boldsymbol{\tau_S} \\ & \wedge \left|\boldsymbol{I_k^H(x,y)} - \boldsymbol{B_k^H(x,y)}\right| \le \boldsymbol{\tau_H} \\ \mathbf{0} & \boldsymbol{sinon} \end{cases} \quad (36)$$

où $\boldsymbol{SP_k}$ est égale à 1 si le point de coordonnées $(\boldsymbol{x},\boldsymbol{y})$ est classé comme ombre, 0 sinon. $\boldsymbol{B_k}$ est l'image de référence (image de fond).

L'explication de l'équation provient de l'observation que lorsqu'une zone est couverte par l'ombre, il en résulte souvent un changement significatif dans la luminance, sans une grande modification dans l'information de couleur. La première condition fonctionne sur la luminance (composante V). Le rapport de luminance a une valeur liée par deux seuils $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ (avec $\mathbf{0} < \boldsymbol{\alpha} < \boldsymbol{\beta} < \mathbf{1}$). L'utilisation de ($\boldsymbol{\beta} < \mathbf{1}$) empêche le système d'identifier, comme ombre, les points où le fond est légèrement obscurci par le bruit, alors que $\boldsymbol{\alpha}$ est utilisé pour définir une valeur maximale de l'effet d'assombrissement causé par l'ombre sur le fond et elle est approximativement proportionnelle à l'intensité de la source lumineuse. Pour la composante S, un seuil pour la différence de saturation entre l'image et la référence est fixé. Les ombres diminuent la saturation des points et expérimentalement, la différence est généralement négative pour les points d'ombre. Pour la composante H, un seuil pour la différence absolue révèle de meilleurs résultats. Cependant, le choix des paramètres $\boldsymbol{\tau_S}$ et $\boldsymbol{\tau_H}$ se fait de manière empirique avec l'hypothèse que la chrominance des points ombrés ou non ombrés varie légèrement (voir la Figure 14).

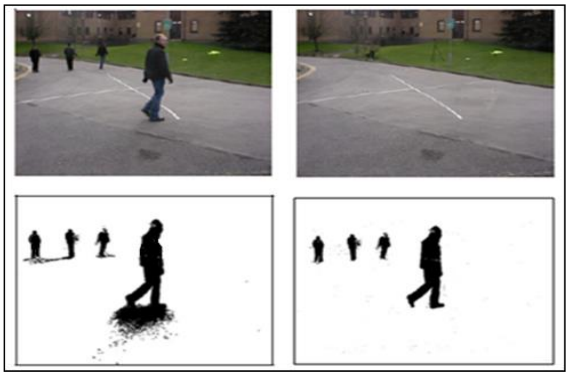

Figure 14. (Haut Gauche) Image originale. (Haut Droite) Arrière-plan estimé à l'instant de l'image. (Bas Gauche) Objets en mouvement extraits par la méthode améliorée du GMM [Zivkovic 2004]. (Bas Droite) Objets en mouvement extraits après suppression d'ombres par la méthode [Cucchiara 2001]

3.2.2. Extraction des masques des blobs

Le masque binaire des objets en mouvement obtenu est affecté par le bruit "poivre et sel" : certains pixels deviennent aléatoirement soit blancs, soit noirs. Afin de réduire l'influence du bruit, une dilatation morphologiques [Najman1994] est appliquée en utilisant comme élément structurant un cercle (la taille du rectangle qui contient l'élément structurant est 3x3). Par la suite, les pixels en mouvement sont regroupés et les régions obtenues (blobs) sont étiquetées à l'aide d'une analyse en composantes connexes proposée dans [Suzuki 1985]. Cette approche exploite la succession, la relation et l'état des pixels voisins les uns des autres pour fournir une liste de contours (liste de points) des régions connexes de l'image binaire. Les blobs de petites tailles sont supprimés. Pour garantir que les masques des objets en mouvement obtenus sont complets (aucun trou n'y existe), une technique de remplissage par diffusion est appliquée [Heckbert 1990]. Elle remplit toute zone de pixels connectés délimitée par des contours de la même couleur. Finalement, l'image de masque binaire des blobs est appliquée sur l'image originale de la scène (une opération ET logique) pour extraire les régions d'intérêt, comme le montre la Figure 15.

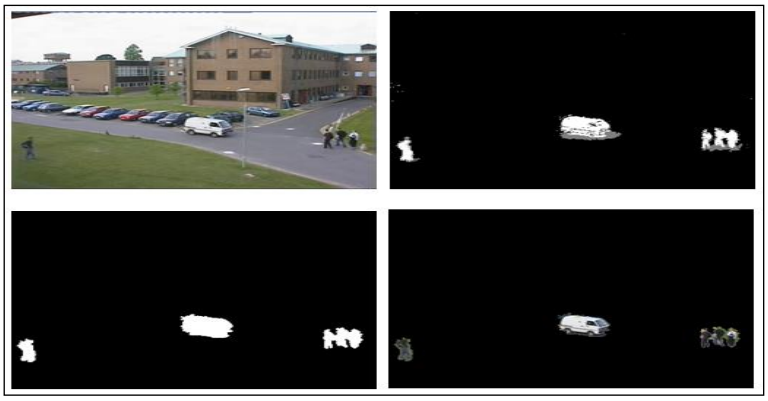

Figure 15. (Haut Gauche) Image originale. (Haut Droite) Image avant-plan : Objets en mouvement en blanc; ombres portées en gris. (Bas Gauche) Masque des blobs extraits. (Bas Droite) Régions d'intérêt extraites

4. Contribution au filtrage spatio-temporel d'objets en mouvement

Généralement, dans les systèmes de vidéosurveillance, l'objectif principal est la compréhension de la scène. Bien que seulement de petites parties du flux vidéo enregistré soient pertinentes, les scènes enregistrées sont entièrement encodées avec une haute qualité et transmises. Dans nos travaux, nous proposons de dégrader la qualité des zones importantes de la scène sans perdre les informations nécessaires pour l'analyse finale. Les régions d'intérêt extraites, dans l'étape précédente (Section 3), contiennent encore des détails; qui sont inutiles et alourdissent les coûts de calcul lors des phases ultérieures (encodage, transmission, analyse). Prenons l'exemple des applications de surveillance routière où les régions d'intérêt

souhaitées sont les véhicules en mouvement : dans ce cas, les informations nécessaires et suffisantes pour analyser le trafic sont la couleur principale, la forme et la trajectoire des objets. Partant de cet exemple, nous proposons une méthode de filtrage spatio-temporel qui simplifie les données extraites afin de ne garder que celles qui sont utiles.

4.1. *Travaux connexes de filtrage vidéo dans les systèmes de vidéosurveillance*

Le filtrage est bien connu comme une technique pour réduire le bruit dans les séquences vidéo et corriger plusieurs types de défauts (rayures, vibrations d'images). Le but principal du filtrage est de supprimer autant d'informations non pertinentes que possible, sans altérer la qualité visuelle des images [Karaca 2000]. Aujourd'hui, les techniques de filtrage sont couramment utilisées dans les travaux de codage vidéo. Par exemple dans [Tsuji 2002], les auteurs proposent un filtre spatio-temporel non linéaire pour lisser le bruit blanc et les textures insignifiantes. Aussi, un schéma de filtrage temporel dans le domaine DCT (Transformée en Cosinus Discrète), décrit dans les travaux de [Byung 2004], est ciblé pour le codage vidéo MPEG (Moving Picture Experts Group). Dans [Cavallaro 2004], les auteurs décomposent la scène en classes contextuelles et classes sémantiques. Les zones appartenant à une classe contextuelle sont soit fixées à une valeur constante ou réduites en importance en utilisant un filtre passe-bas pour améliorer les performances des codeurs vidéo. Citons également les travaux de [Ziliani 2003], qui proposent un filtre temporel en se basant sur les résultats du suivi de l'objet pour améliorer les performances des codeurs dans le contexte de la vidéosurveillance. Les zones mobiles sont détectées et suivies d'une image à une autre pour comprendre leurs propriétés temporelles. Toutes les informations visuelles sans importance et les objets mobiles incohérents sont supprimés, tandis que les régions d'intérêt sont maintenues dans leur qualité originale ce qui minimise les coûts de codage. La limitation du travail décrit dans [Cavallaro 2004] est à définir les modalités du filtrage sans tenir compte de deux critères principaux : le type de l'objet et les spécificités du processus de codage. La solution dans [Ziliani 2003] est intéressante, mais le système proposé ne correspond pas aux exigences en termes de temps et des coûts de calcul des applications de surveillance en temps réel. L'intégration de la tâche de suivi d'objet au niveau de la caméra présente des charges supplémentaires qui ne peuvent pas souvent être accomplies. De plus, en vertu de cette approche, l'étape de suivi est exécutée dans deux occasions : une première fois au niveau de l'acquisition pour le filtrage et ensuite au niveau de la réception pour l'analyse. L'utilisation du

suivi d'objets pour assurer une bonne décomposition en régions significatives et non significatives peut être remplacée par une technique fiable de séparation de fond.

4.2. *Généralités sur le processus de codage vidéo*

Pendant la phase de codage vidéo, les prédictions intra-trames et inter-trames sont les clés pour avoir une haute qualité de compression. Pour être codée, une image est divisée en blocs appelés macro-blocs. Chaque macro-bloc est prédit à partir de son voisinage (dans le cas de la prédiction intra-trame), ou à partir des images de référence (dans le cas de la prédiction inter-trame). Les images de référence connues sous le nom de trame I, sont codées uniquement à partir de leur contenu avec un algorithme de prédiction spatiale et n'ont pas besoin de données supplémentaires pour être décodées. Le codage inter-trame, pour les images prédites (P), consiste à trouver le bloc le plus similaire au bloc courant dans l'image de référence grâce à un algorithme de block-matching (correspondance des blocs) [Barjatya 2004]. Si la recherche est réussie, le bloc est codé par un vecteur, connu sous le nom de vecteur de mouvement, qui indique la position du bloc correspondant dans l'image de référence. Dans la plupart des cas, le bloc trouvé n'est pas une correspondance exacte. C'est pourquoi, le codeur calcule la différence entre les deux blocs (voir Figure 16). Ces valeurs résiduelles, appelées les erreurs de prédiction, doivent être transformées et envoyées vers le décodeur. Si l'algorithme de prédiction est en mesure de trouver un bloc avec des valeurs très proches de celles du bloc courant, l'erreur de prédiction devient plus petite. Alors, une fois transformée et compressée, la taille du vecteur de mouvement et du bloc résiduel ensemble sera inférieure au bloc courant non compressée. Au contraire, si la prédiction n'arrive pas à trouver un bloc semblable, l'erreur de prédiction sera plus grande et le flux codé prendra une taille plus importante que le bloc courant non compressé.

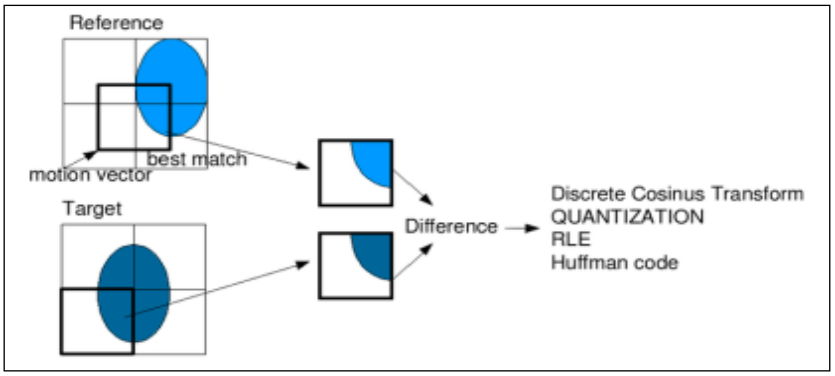

Figure 16. Processus de prédiction inter-trame : la différence entre le bloc dans l'image de référence et son bloc similaire est l'erreur de prédiction de ce bloc

4.3. *Schéma de l'approche proposée*

La méthode que nous présentons dans cette partie est différente, même si elle repose sur des hypothèses similaires des travaux de [Ziliani 2003]. La séquence enregistrée est simplifiée selon une définition précise de ce que sont les informations vidéo importantes. Pour démarrer le filtrage, les parties pertinentes des informations visuelles doivent être séparées des informations contextuelles. Définir la partition sémantique dépend de la tâche analytique à effectuer. Par conséquent, certaines connaissances a priori sur l'objet à segmenter sont nécessaires. Dans le contexte de la vidéosurveillance, les informations de mouvement forment les informations sémantiques pour segmenter les objets en mouvement. Ainsi, les zones portant des objets mobiles représentent les informations demandées. Bien qu'elles correspondent à de petites portions de la vidéo enregistrée, ces parties contiennent encore des informations supplémentaires qui ne sont pas nécessaires et alourdissent les coûts de calcul. La méthode de simplification proposée filtre les parties de la scène qui ne seront pas utilisées au cours de l'étape d'analyse. La suppression de l'information visuelle inutile fournit des caractéristiques d'intérêt équivalentes tout en réduisant la qualité et les coûts de codage. Tout d'abord, les informations pertinentes (avant-plan) sont séparées des informations contextuelles (fond). Ensuite, les informations sont simplifiées en supprimant les parties de la scène qui ne sont pas intéressantes ce qui réduit le taux d'informations à traiter. Ce processus est mis en œuvre en tant que la phase de pré-analyse, suivie de l'étape de codage vidéo. Pendant le processus de codage, le coût est fortement déterminé par l'erreur de prédiction entre les macro-blocs. Si elle est petite, le taux de compression est élevé. La méthode proposée tient compte des spécificités du processus de codage : la simplification des données se fait de manière à ce que deux blocs mis en correspondance dans deux images successives aient presque la même valeur des composantes couleurs. Ainsi, l'erreur de prédiction devient minimale, le coût de codage devient plus faible et la compression est meilleure.

Les régions contenant des objets en mouvement sont extraites suivant les étapes décrites dans la Section 3. La séquence vidéo résultante (contenant seulement les régions extraites d'intérêt) est divisée en groupe d'images (GOP). Chaque première image (image de référence) d'un GOP représentera la trame codée (I) d'un GOP de la séquence codée. Les images restantes du GOP formeront les trames prédites (P) du GOP de la séquence codée (illustré sur la Figure 17). Nous supposons que les GOP de la séquence vidéo codée ne contiennent qu'une unique trame I et plusieurs trames P. L'Algorithme 1 décrit le processus de filtrage d'un GOP. La méthode fonctionne d'une manière spatio-temporelle. Elle consiste à

appliquer, pour chaque GOP, un filtrage spatial sur la trame I, puis un filtrage temporel sur chaque image P. Le filtrage est réalisé de façon que seule les informations pertinentes de la séquence sont conservées pour l'analyse. Également, les redondances spatiales et temporelles sont incrémentées pour de meilleures performances de codage. Il convient de noter que les images simplifiées obtenues (visualisées à la réception) sont formées par l'image de l'arrière-plan filtrée spatialement superposée par les régions spatio-temporellement filtrées.

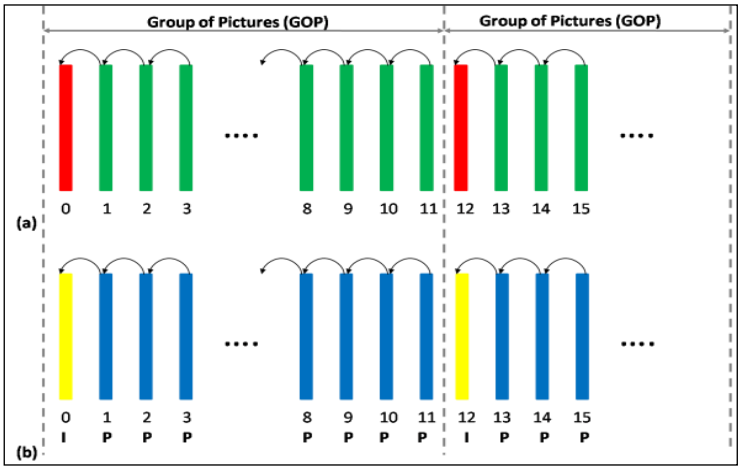

Figure 17. La distribution d'images dans un groupe d'images (taille du GOP = 12) dans (a) la séquence simplifiée et (b) la séquence codée. (Rouge) image filtré spatialement; (Vert) image filtré temporellement; (Jaune) image prédite spatialement; (Bleu) image prédite temporellement

Algorithme 1: **Filtrage spatio-temporel de la séquence vidéo**

```
1: for each GOP (I 1..N) do
2:         filtrage spatial (I)
3:         for F from 1 to N do
4:                 filtrage temporel (F)
5:         end for
6: end for
```

4.3.1. Filtrage spatial

Le filtrage spatial est une simplification intra-trame appliquée sur l'image de référence $\boldsymbol{I}$. Une fenêtre glissante passe à travers l'image. A chaque pas, la moyenne robuste des valeurs des pixels de la fenêtre est calculée pour remplacer la valeur du pixel central de la fenêtre. L'Algorithme 2 explique le processus du filtrage spatial avec : (1) $\boldsymbol{SW}$ est la fenêtre glissante de largeur $\boldsymbol{W_{SW}}$ et hauteur $\boldsymbol{H_{SW}}$, $\boldsymbol{S_{SW}}$ est le pas de glissement, $\mathbf{P_{SW}}$ est la valeur du pixel central de la fenêtre glissante $\boldsymbol{SW}$; (2) $\boldsymbol{W}$ et $\boldsymbol{H}$ sont respectivement la largeur et la hauteur de l'image; (3) $\boldsymbol{M}$ est la valeur de la moyenne robuste calculée. C'est un vecteur à trois éléments

représentant les valeurs des composantes $\{\boldsymbol{R}, \boldsymbol{G}, \boldsymbol{B}\}$ du pixel; (4) $\boldsymbol{H_C}$ est l'histogramme couleur de la composante $\boldsymbol{C}$, $\boldsymbol{T}$ est une valeur en pourcentage.

Algorithme 2 : **Filtrage spatial**

1: **for** i from 0 to W **do**
2: **for** j from 0 to H **do**
3: M= moyenne robuste ($\boldsymbol{SW}$ $(\boldsymbol{i}, \boldsymbol{j})$)
4: $\boldsymbol{P_{SW}}$=M
5: j=j+$\boldsymbol{S_{SW}}$
6: **end for**
7: i=i+$\boldsymbol{S_{SW}}$
8: **end for**

L'Algorithme 3 explique comment calculer la moyenne robuste à chaque itération. La moyenne robuste est la moyenne des T premières principales couleurs dans l'histogramme des couleurs $\{\boldsymbol{R}, \boldsymbol{G}, \boldsymbol{B}\}$ d'une fenêtre glissante (la valeur de $\boldsymbol{T}$ est fixée expérimentalement). L'utilisation de telle moyenne permet l'obtention d'une valeur de couleur représentative des couleurs dominantes dans la fenêtre. De cette façon, en filtrant spatialement l'image, l'information de couleurs est réduite à la moyenne des couleurs dominantes.

Algorithme 3 : **Moyenne robuste**

1: **for each** composante couleur C **in** {R,G,B} **do**
2: $\boldsymbol{H_C}$= calcul de l'histogramme couleurs des pixels de la fenêtre
3: $\boldsymbol{M_C}$= calcul de la moyenne des $\boldsymbol{T}$ premières couleurs dominantes dans l'histogramme $\boldsymbol{H_C}$
4: **end for**

4.3.2. Filtrage temporel

De façon similaire au filtrage spatial, le filtrage temporel est une simplification inter-trame appliquée sur les images restantes du GOP en se basant sur l'image de référence déjà filtrée. Les blocs de l'image sont modifiés pour devenir semblables à leurs correspondants dans l'image de référence. Pour chaque région d'intérêt $\boldsymbol{ROI}$ dans l'image $\mathbf{F_N}$ du GOP, nous recherchons sa correspondante $\boldsymbol{ROI_{BM}}$ dans l'image précédente $\mathbf{F_{N-1}}$ en utilisant l'algorithme de block-matching. Un score de similarité $\boldsymbol{R}$ (Equation 37) basé sur la méthode des différences quadratiques normalisées est calculé [Rodgers 1988]. Ainsi, pour une similarité parfaite $\boldsymbol{R}$ est égale à $\mathbf{0}$ et pour une mauvaise correspondance, la valeur de $\boldsymbol{R}$ est grande. Si la recherche est réussie ($\boldsymbol{R} < \boldsymbol{T_{BM}}$, $\boldsymbol{T_{BM}}$ est un seuil fixé expérimentalement), chaque pixel obtient la même valeur que son apparié déjà simplifié; sinon, la région est filtrée spatialement. Le processus appliqué pour chaque image $\boldsymbol{F}$ est exprimé dans l'Algorithme 4.

$$\boldsymbol{R(x,y)} = \frac{\sum_{x',y'}[\boldsymbol{ROI}(x',y') - \boldsymbol{F_{N-1}}(x+x',y+y')]^2}{\sqrt{\sum_{x',y'} \boldsymbol{ROI}(x',y')^2 . \sum_{x',y'} \boldsymbol{F_{N-1}}(x+x',y+y')^2}} \quad (37)$$

Algorithme 4 : **Filtrage temporel**

```
1: for each ROI in F_N do
2:     R, ROI_BM = recherche de la région similaire de ROI dans F_{N-1} par block-matching
3:     if R < T_BM then
4:         for i from 0 to W_ROI do
5:             for j from 0 to H_ROI do
6:                 ROI(i,j) = ROI_BM(i,j)
7:             end for
8:         end for
9:     else
10:        Filtrage spatial (ROI)
11:    end if
12: end for
```

4.3.3. Encodage

Les vecteurs de mouvement calculés au cours du block-matching pour le filtrage temporel sont réutilisés lors de la phase d'encodage. Les valeurs des erreurs de prédiction sont nulles pour les blocs filtrés temporellement et réduites pour ceux filtrés spatialement. De cette façon, l'erreur de prédiction est généralement très faible. Par conséquent, l'encodage est plus rapidement exécuté (coûts de calcul inférieurs) avec une meilleure compression.

Grâce à cette méthode, les informations non pertinentes sont modifiées ou éliminées d'une manière intelligente prenant en considération les caractéristiques nécessaires pour l'analyse finale. Tandis que lors de l'encodage dans les systèmes de surveillance habituels, les données sont modifiées ou supprimées d'une façon aveugle. À travers le filtrage spatio-temporel, de suffisantes informations sont maintenues pour le suivi d'objets (la position de l'objet à chaque image) et la classification d'objets (la couleur dominante, la forme, la taille). Ces caractéristiques sont suffisantes pour les applications de vidéosurveillance du premier et du deuxième niveau, comme la plupart des systèmes d'analyse du trafic routier. Les cas spécifiques d'identification des descripteurs uniques tels que les plaques d'immatriculation, les visages, ... ne sont pas abordés dans nos travaux. Nous sommes d'accord qu'ils doivent être précisément détectés et codés, mais c'est un problème qui dépasse le contexte de cette thèse.

5. Contribution à la modélisation géométrique d'objets en mouvement

Parmi les problèmes rencontrés lors du suivi et de la classification d'objets est les groupes d'objets en occlusion. Simplifier la représentation des objets mène à un suivi d'objets plus fiable avec moins d'informations utilisées mais une préservation des caractéristiques nécessaires. Par conséquent, la modélisation d'objets en mouvement en une forme plus simple peut être considérée comme une technique de pré-analyse. Les objets peuvent être représentés de différentes façons et le choix de la représentation d'un objet dépend fortement du domaine d'application.

5.1. *Travaux connexes de modélisation d'objets dans les systèmes de vidéosurveillance*

La représentation de l'objet est un choix crucial dans la compréhension de la scène, car il intègre les caractéristiques qui sont nécessaires durant les étapes suivantes. Dans la littérature, de nombreux travaux modélisent les objets dans différentes simples formes 2D. Comme [Cucchiara 2005] où les objets sont représentés comme rectangles, [Comaniciu 2003] comme des ellipses. Ces modèles peuvent être rapidement calculés. Ils sont utilisés dans plusieurs applications où les deux dimensions de l'objet sont satisfaisantes pour analyser la scène : la détection d'arrêt des véhicules dans les routes [Melli 2005], la détection de la posture humaine [Cucchiara 2005], le suivi des groupes de personnes dans un train [Cupillard 2001], etc. Dans ces applications citées, le modèle 2D est assez suffisant pour trouver la position 3D de l'objet. Également, le calcul du modèle 2D est peu coûteux et répond à la contrainte du temps réel. Cependant, ils sont imprécis car ils dépendent de la déformation de l'objet, de ses positions et de ses rotations relatives par rapport à la caméra. Dans l'autre extrême, nous trouvons les modèles d'objets spécifiques (véhicules, piétons). Ils sont très dépendants de l'objet et de l'application tels que les modèles articulés dans [Boulay 2006]. Ces modèles sont très précis et conduisent à de bons résultats de détection. Toutefois, ils ne sont pas adéquats pour les applications en temps réel (coûts de calcul élevés) et manquent de flexibilité pour représenter des objets en général. Au milieu, pour faire face aux limites des modèles cités précédemment, les modèles 3D ont été développés avec différentes formes : cylindres en [Scotti 2005], parallélépipèdes en [Yoneyama 2005], etc. Ils sont peu coûteux et appropriés pour les applications en temps réel. Ils peuvent représenter divers objets avec le maintien des caractéristiques pertinentes, mais ils ont besoin d'avoir certaines connaissances préalables telles que : le modèle prédéfini, les paramètres intrinsèques de la caméra, etc.

Comme exemple : les auteurs de [Lee 2009] proposent un simple modèle planaire 3D, se basant sur la projection 2D de l'objet. La forme plane reste une représentation très vague. Dans [Zuniga 2006, Zuniga 2011], Zuniga et al. travaillent sur l'estimation du modèle parallélépipédique 3D. Ils utilisent un ensemble de régions mobiles 2D, la matrice de la transformée perspective (obtenu depuis la calibration de la caméra) et un modèle prédéfini 3D des objets attendus dans la scène afin de trouver le modèle 3D le plus adéquat à l'objet.

5.2. *Schéma de l'approche proposée*

La méthode que nous proposons fait face aux inconvénients déjà cités en estimant un modèle parallélépipédique des objets dans le plan de l'image. Elle est purement géométrique et développée pour les applications mono-caméra sans aucune connaissance préalable sur l'orientation de l'objet et sur les paramètres intrinsèques de la caméra et de la scène. Cette technique représente l'objet en mouvement (principalement les véhicules) dans une forme parallélépipédique afin de simplifier sa représentation. En même temps, les caractéristiques requises pour le suivi et la classification d'objets sont maintenues.

Pour simplifier la phase d'identification des véhicules, nous supposons que (voir la Figure 18):

1) L'axe de la caméra est parallèle à la route, et l'appareil lui-même est orienté verticalement (pour maintenir l'orientation verticale après la projection sur le plan de l'image).

2) Chaque véhicule est modélisé sous la forme d'un parallélépipède. Après sa projection sur le plan de l'image, ce modèle est défini par les points A, B, C, D, E et F (comme illustré sur la Figure 18) avec :

- Les segments AF et CD sont parallèles à la ligne de circulation.
- Les segments AB et DE sont horizontaux.
- Les segments BC et EF sont verticaux dans le plan de l'image.

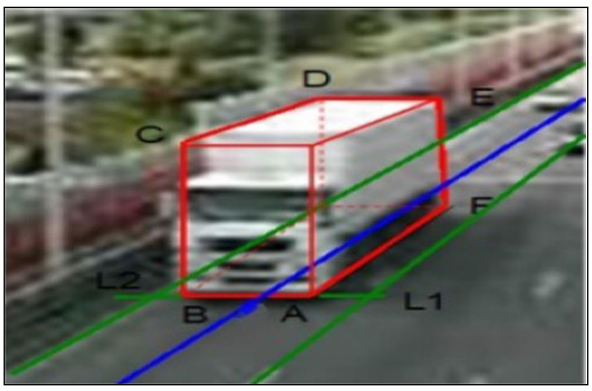

Figure 18. Représentation parallélépipédique d'un objet en mouvement

5.2.1. Construction du modèle parallélépipédique

Après l'étape d'extraction des régions d'intérêt (Section 3), les objets en mouvement sont détectés en tant que blobs pour calculer leurs modèles parallélépipédiques. L'idée de cette représentation est de trouver un parallélépipède délimité par les bords de la boîte englobante. Pour chaque objet en mouvement $\boldsymbol{O}$, sa boite englobante $(\boldsymbol{A_1A_2A_3A_4})$ est calculée. L'étape suivante consiste à extraire les contours externes $\boldsymbol{ctr}$ de l'objet (Figure 19.a). Ensuite, pour chaque sommet $\boldsymbol{A_{1\leq i\leq 4}}$, nous recherchons le point le plus proche appartenant au contours externes $\boldsymbol{P_{1\leq i\leq 4}}$ avec :

$$\forall \boldsymbol{A_{1\leq i\leq 4}}, \exists\, \boldsymbol{P_{1\leq i\leq 4}} \,/\, \boldsymbol{d}(\boldsymbol{A_i}, \boldsymbol{P_i}) = \mathbf{argmin}\, \boldsymbol{d}(\boldsymbol{A_i}, \boldsymbol{C}) \qquad (38)$$

où $\boldsymbol{d}(.,.)$ est la distance euclidienne, $\boldsymbol{C} = \{\forall\, \boldsymbol{point}\, \boldsymbol{P} \in \boldsymbol{ctr}\}$.

A ce stade, pour chaque sommet $\boldsymbol{A_i}$, le point le plus proche $\boldsymbol{P_i}$ est sélectionné. Nous cherchons le triangle rectangle ayant comme sommet rectangle $\boldsymbol{A_i}$ et hauteur $(\boldsymbol{d})$ (passant par $\boldsymbol{A_i}$ et $\boldsymbol{P_i}$). Cela revient à trouver la droite $(\boldsymbol{d'})$ perpendiculaire à $(\boldsymbol{d})$ en se basant sur le système d'équations suivant :

$$\begin{cases} (\boldsymbol{d}): \boldsymbol{a_1}\mathbf{x} + \boldsymbol{b_1} = \mathbf{y} \\ (\boldsymbol{d'}): \boldsymbol{a_2}\mathbf{x} + \boldsymbol{b_2} = \mathbf{y} \\ (\boldsymbol{d}) \perp (\boldsymbol{d'}) \end{cases} \Leftrightarrow \begin{cases} (\boldsymbol{d}): \boldsymbol{a_1}\mathbf{x} + \boldsymbol{b_1} = \mathbf{y} \\ (\boldsymbol{d'}): \boldsymbol{a_2}\mathbf{x} + \boldsymbol{b_2} = \mathbf{y} \\ \boldsymbol{a_1} * \boldsymbol{a_2} = -\mathbf{1} \end{cases} \qquad (39)$$

Puisque les coordonnées de $\boldsymbol{A_i}$ et $\boldsymbol{P_i}$ sont déjà connues, la droite $(\boldsymbol{d'})$ est obtenue en résolvant le système d'équations (39). Ainsi, comme montré dans la Figure 19.b, le triangle rectangle de chaque sommet est construit. Pour chaque deux sommets opposés de la boite englobante, le triangle avec la surface minimale est maintenu. Ensuite, le plus grand parmi ces deux triangles est gardé (Figure 19.c). De cette manière, nous garantissons que l'objet est entièrement inclus dans sa forme parallélépipédique. Enfin, pour obtenir le modèle final, la surface du triangle gardé (triangle vert de sommet $\boldsymbol{A_4}$ dans la Figure 19.c) est soustraite des deux sommets opposés comme illustré sur la Figure 19.d. Le modèle $\boldsymbol{M}(\boldsymbol{O})$ obtenu répond aux hypothèses précédemment posées (Figure 19.e et Figure 19.f).

5.2.1. Affinement du modèle parallélépipédique

A cause des scénarios de collisions d'objets ou d'objets trop proches, au cours de la phase de segmentation plusieurs objets sont associés à une même région d'intérêt. Ce type d'erreurs de segmentation perturbe la phase de modélisation puisque plusieurs objets seront

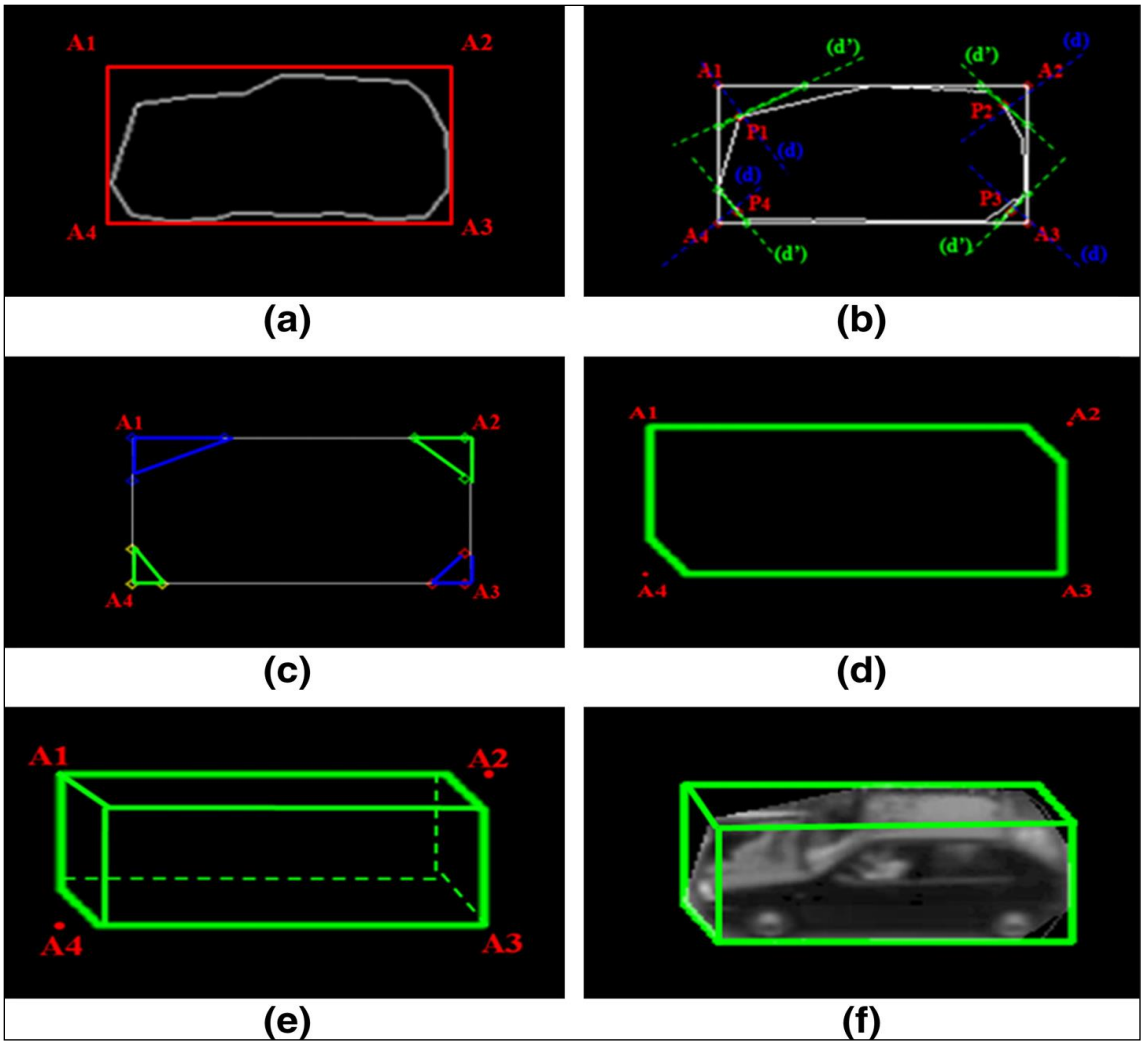

Figure 19. Les étapes de modélisation parallélépipédique d'objet

représentés par un seul modèle. Cet inconvénient est évité en imposant au modèle $\boldsymbol{M(O)}$ de maximiser la mesure suivante :

$$\boldsymbol{M(O) = argmax_M}\left(\boldsymbol{CARD[M]} \,/\, \frac{\boldsymbol{CARD[M \cap O]}}{\boldsymbol{CARD[M]}} > \alpha\right) \quad \textbf{(40)}$$

Elle permet de calculer le pourcentage de la zone couverte par le modèle par rapport à la surface de tout l'objet, avec $\boldsymbol{\alpha}$ est un coefficient introduit pour tenir compte des erreurs de segmentation et le fait que le modèle est seulement une approximation de la réalité (généralement $\boldsymbol{\alpha}$=0.9).

Au cas où la valeur de $\boldsymbol{M(O)}$ serait inférieure à $\boldsymbol{\alpha}$, le modèle est affiné en lançant une recherche exhaustive qui tient compte des propriétés du modèle (l'orientation et le parallélisme des segments du modèle déjà décrits dans les hypothèses). La taille du modèle obtenu est réduite dans toutes les directions. À chaque itération, le modèle avec la valeur maximale de $\boldsymbol{M(O)} > \alpha$ est maintenu (voir Figure 20). Le reste de la région $\boldsymbol{O}$ est modélisé

par un autre modèle parallélépipédique de la même façon afin d'obtenir le meilleur modèle clôturant exactement toute la région (voir Figure 21).

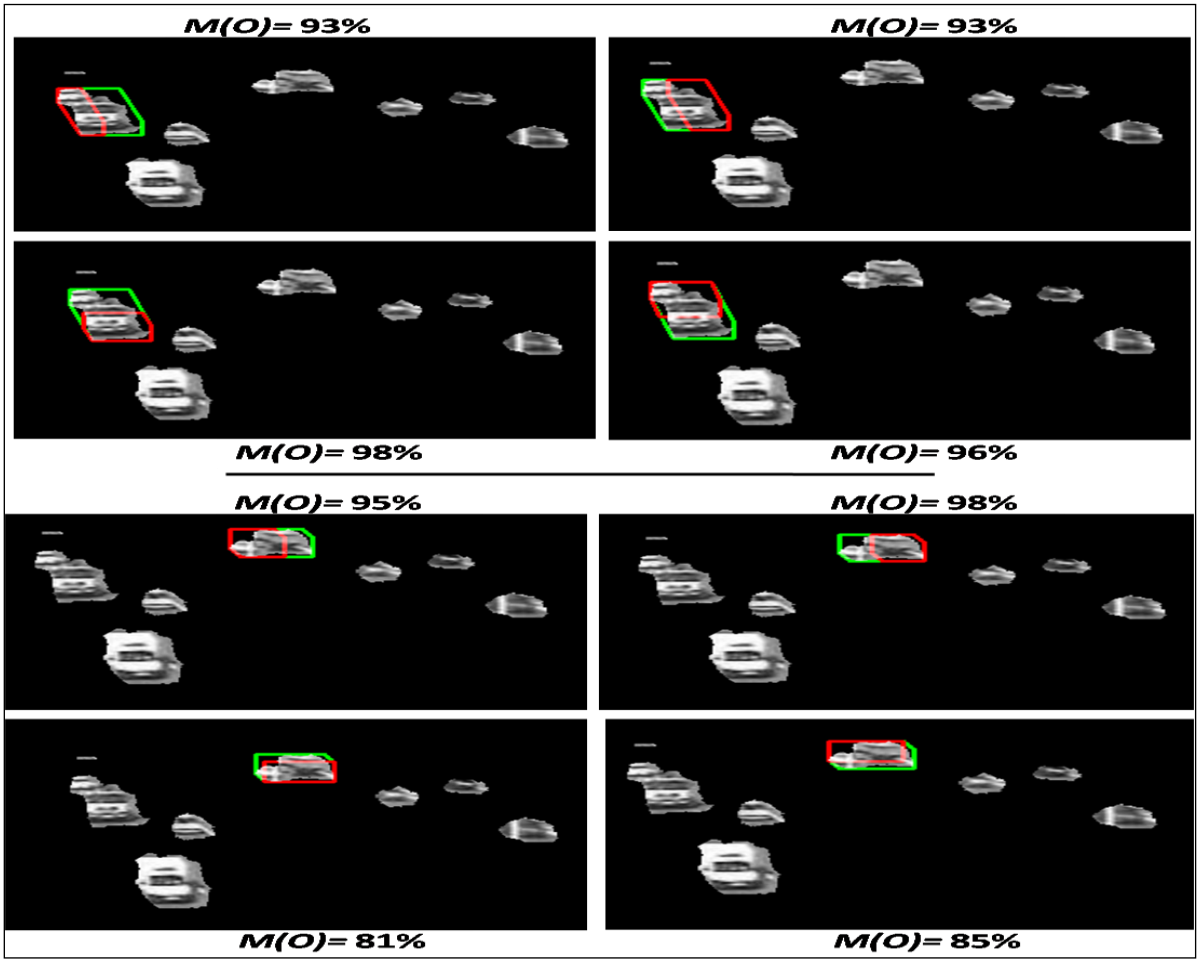

Figure 20. Les étapes d'affinement du modèle parallélépipédique dans les cas de collusions d'objets ou d'objets trop proches

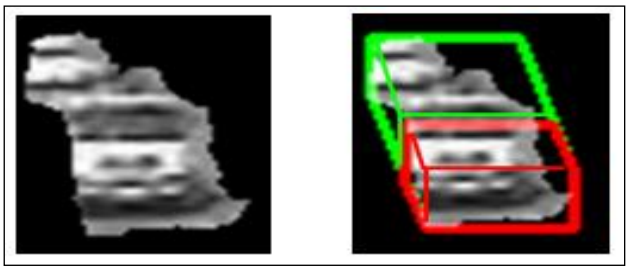

Figure 21. Exemple de résultat de la modélisation parallélépipédique après affinement

La méthode décrite est indépendante de la nature de l'objet, elle peut modéliser une variété d'objets (véhicules, piétons, ...). Elle est adéquate pour les applications en temps réel grâce à son coût de calcul faible. Le modèle obtenu permet de déterminer facilement la classe de l'objet en calculant le rapport hauteur/largeur qui devrait être plus élevé pour les piétons que pour les véhicules. Également, il permet d'obtenir une bonne approximation des

dimensions et de la position de l'objet dans la scène. L'orientation du modèle mène à définir la direction principale du déplacement de l'objet. Cette représentation est indépendante des paramètres de la caméra et de l'objet.

6. Résultats expérimentaux

6.1. *Configuration*

Toutes les expérimentations sont effectuées sur un ordinateur Windows7-64bits équipé d'un processeur Intel i7-2670QM CPU @ 2.20GHz, avec 4.0GB de mémoire RAM. L'évaluation utilise des scènes de surveillance de trafic routier et de lieux publics. Certaines séquences de tests appartiennent à des bases de données connues de vidéosurveillance : VISOR[1], PETS[2] et I-LIDS[3]. D'autres sont des scènes autoroutières fournies par la société Adacis[4]. Elles sont toutes issues d'acquisition dans des conditions réelles en utilisant une caméra fixe avec différentes positions et angles (voir Figure 22). Les zones contenant des objets mobiles tels que des piétons, des véhicules représentent les régions d'intérêt requises.

Figure 22. Exemple d'images extraites des séquences de test

6.2. *Outils d'évaluation*

Les contributions proposées sont développées avec le langage orienté objet C++ en intégrant la bibliothèque OpenCV pour le traitement des images et des vidéos. La

1 http://www.openvisor.org

2 http://www.cvg.rdg.ac.uk/PETS2009/a.html

3 https://www.gov.uk/imagery-library-for-intelligent-detection-systems

4 http://www.adacis.net

compression vidéo H.264/AVC [Wedi 2003] est réalisée par le logiciel de référence H.264/14496-10 AVC[5].

Pour l'évaluation applicative des méthodes suggérées, les tâches analytiques de comptage, suivi et analyse comportementale d'objets sont effectuées par le système de vidéosurveillance de la bibliothèque OpenCV [Chen 2005]. Le schéma général du système est présenté dans la Figure 23. Le premier module "FG/BG Detection" effectue la classification avant-plan/arrière-plan des pixels de l'image en utilisant la méthode de Li [Li 2003]. Cette technique se base sur les statistiques des couleurs des pixels et des co-occurrences des couleurs. Les distributions des couleurs des pixels et des co-occurrences des couleurs sont représentées par des histogrammes. Des règles de décisions Bayésiennes sont appliquées pour classer le pixel en avant-plan ou arrière-plan. Le module "Blob Entering Detection" utilise le résultat du premier module pour détecter un nouvel objet qui pénètre dans la scène. Il se base sur le suivi des composantes connexes en calculant et détectant les blobs dans les images successives [Senior 2006]. Le troisième module réalise un suivi image par image de la position et de la taille du blob. C'est un suivi d'objets hybride qui combine deux composants : un suivi des composantes connexes qui fournit des résultats de suivi fiables et rapides quand il n'y a pas de collusions d'objets et un autre suivi basé sur les algorithmes de meanshift et de filtrage à particules [Comaniciu 2000, Nummiaro 2003]. Un filtre de Kalman [Kalman 1960] est utilisé pour prédire la position de l'objet dans l'image suivante. Si une collusion va se produire, le suivi avec le filtre à particules est appliqué, sinon, le suivi des composantes connexes est utilisé. Les deux modules suivants "Trajectory PostProcessing" et "Trajectory Generation" servent, respectivement, pour lisser les trajectoires et les enregistrer. Le dernier module analyse les trajectoires et détecte les anomalies. Une approche qui traite une trajectoire comme un ensemble indépendant de vecteurs de caractéristiques (la position du blob, la vitesse du blob et la durée de l'état du blob) est adoptée. Un histogramme 5D de ces caractéristiques est continuellement collecté et analysé. Ainsi, si le blob actuel a des caractéristiques qui ne sont jamais ou rarement observés avant, alors le blob et ses trajectoires sont classés comme anormaux.

Quant à la classification d'objets, la méthode proposée dans les travaux de Zang [Zang 2003] est utilisée pour la séparation des piétons et des véhicules (voir la Figure 24). Elle utilise le rapport hauteur/largeur de la boite englobante de l'objet en mouvement. Pour un véhicule, cette valeur doit être inférieur à 1, pour un piéton, cette valeur doit être supérieure à

5 http://iphome.hhi.de/suehring/tml/

1.5. Pour tenir compte des situations particulières où le rapport est compris entre 1et 1.5, les coins de l'objet sont extraits à l'aide du détecteur des coins SUSAN pour classer l'objet comme un véhicule ou un piéton (un véhicule produit plus de coins).

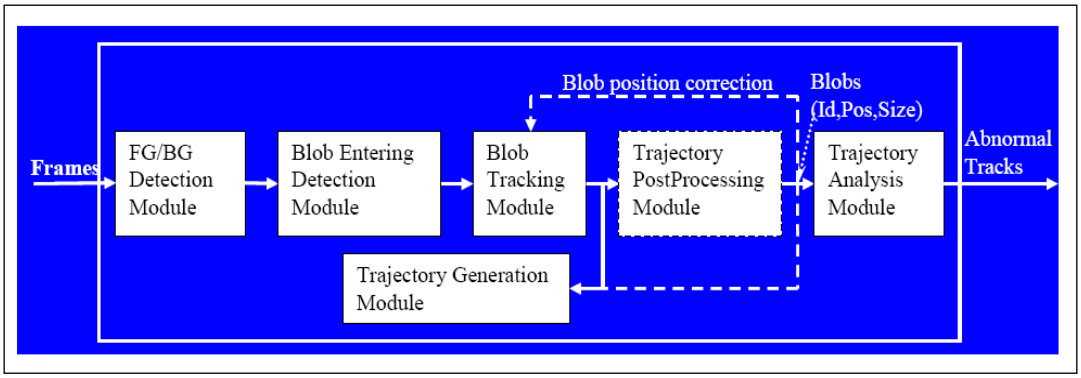

Figure 23. Schéma du système de vidéosurveillance OpenCV [Chen 2005]

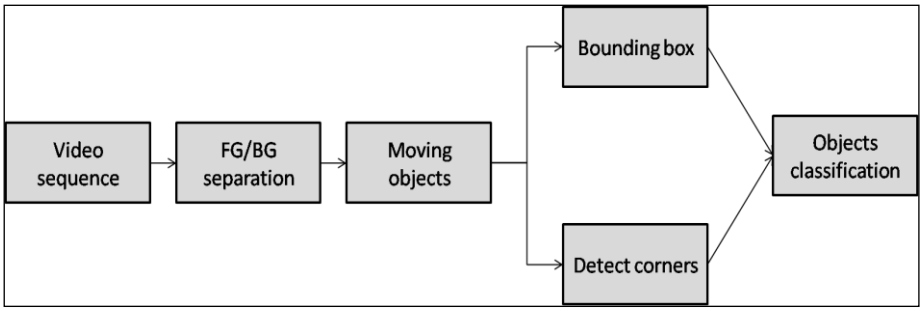

Figure 24. Schéma de la méthode de Zang pour la classification d'objets [Zang 2003]

6.3. *Métriques d'évaluation*

L'évaluation perceptive nécessite des mesures de dégradation de la qualité d'image. Cela peut être fait avec des mesures objectives qui devraient refléter les préférences des yeux humains. Ici, nous choisissons d'utiliser les mesures objectives les plus populaires : le rapport signal sur bruit (***PSNR***) et l'indice de similarité (***SSIM***) [Zhou 2004]. Le ***PSNR*** mesure la qualité de reconstruction d'image par rapport à l'originale. Il est mesuré par le rapport entre la valeur maximale possible de luminance d'un pixel de l'image x_{max} et les erreurs introduites par la reconstruction :

$$PSNR = 10log_{10}\frac{(x_{max})^2}{MSE} \qquad (41)$$

Où ***MSE*** est l'erreur quadratique moyenne définie entre deux images I_0 et I_r de taille $m * n$:

$$\boldsymbol{MSE = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I_0(i,j) - I_r(i,j))^2} \qquad (42)$$

L'idée du $\boldsymbol{SSIM}$ est de mesurer la similarité structurelle entre deux images. La mesure comprend des modèles perceptifs implicites qui reflètent les caractéristiques du système visuel humain. Elle est définie par les statistiques locales comparant le contraste, la luminance, et la structure. Etant donné deux échantillons $\boldsymbol{x}$ et $\boldsymbol{y}$ extraits de la même position spatiale de deux images, la métrique SSIM est calculée comme suit :

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \qquad (43)$$

où $\boldsymbol{\mu_x}$ et $\boldsymbol{\mu_y}$ sont les intensités moyennes, $\boldsymbol{\sigma_x^2}$ et $\boldsymbol{\sigma_y^2}$ sont les variances, $\boldsymbol{\sigma_{xy}}$ est la covariance de $\boldsymbol{x}$ et $\boldsymbol{y}$, et $\boldsymbol{C_1}$ et $\boldsymbol{C_2}$ sont des constantes de faibles valeurs utilisées pour stabiliser la division avec un faible dénominateur. Les valeurs de $\boldsymbol{SSIM}$ varient entre $\boldsymbol{0}$ et $\boldsymbol{1}$, et la valeur $\boldsymbol{1}$ est atteinte seulement dans le cas de deux images identiques.

Les métriques utilisées pour évaluer les performances des méthodes sont les mesures de précision et de rappel.

$$\boldsymbol{Précision = \frac{TP}{TP+FP}} \qquad (44)$$

$$\boldsymbol{Rappel = \frac{TP}{TP+FN}} \qquad (45)$$

avec : $\boldsymbol{TP}$ (vrais positifs, présence d'un objet correctement détecté), $\boldsymbol{FP}$ (faux positifs, détection non présente dans la vérité-terrain) et $\boldsymbol{FN}$ (faux négatifs, présence d'un objet non détecté).

6.4. *Evaluation de l'extraction des régions d'intérêt*

L'étape d'extraction des régions d'intérêt est réalisée à travers l'approche de [Zivkovic 2004] pour l'extraction de fond suivie de la suppression des ombres suivant le travail de [Cucchiara 2001]. Dans la Figure 25, cette méthode utilisée pour l'extraction des régions d'intérêt est évaluée par rapport à la GMM originale de [Stauffer 1999]. Nous comparons les résultats avec la vérité terrain, qui consiste en une extraction manuelle des objets dans chaque image de la scène. Les valeurs de rappel et de précision obtenues sont meilleures que [Stauffer 1999] (une augmentation de 2% à 20% est observée). Ils varient entre 0,78 et 0,98 en fonction de la complexité de la scène.

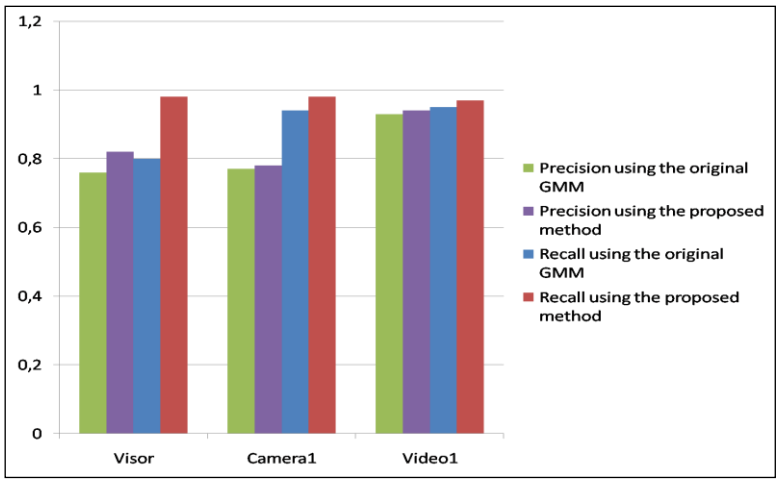

Figure 25. Valeurs de rappel et de précision calculées pour l'étape d'extraction des régions d'intérêt dans différentes séquences en utilisant la méthode des GMM originale [Stauffer 1999] et la méthode proposée [Zivkovic 2004]

6.5. *Evaluation du filtrage spatio-temporel d'objets en mouvement*

Pour l'évaluation des performances, les paramètres du filtrage spatio-temporel ont les valeurs suivantes : le $\boldsymbol{GOP}$ est de taille $\mathbf{12}$ images, la fenêtre glissante $\boldsymbol{SW}$ est de $\mathbf{7} * \mathbf{7}$ pixels, le pas de glissement $\boldsymbol{S_{SW}}$ est $\mathbf{1}$ pixel, la valeur de $\boldsymbol{T}$ est $\mathbf{60}\%$ et la valeur de $\boldsymbol{T_{BM}}$ est 0.5. La taille de la fenêtre $\boldsymbol{SW}$ est, aussi, fixée empiriquement car des valeurs plus élevées (au delà de $\mathbf{7} * \mathbf{7}$) conduisent à une qualité visuelle dégradée des régions extraites (essentiellement les régions représentant des piétons).

6.5.1. Évaluation perceptive

La qualité des images obtenues est évaluée en utilisant des méthodes subjectives et objectives avant et après l'encodage H.264/AVC. Depuis que le filtrage spatio-temporel réduit la qualité de l'image en éliminant l'information visuelle inutile, nous examinons la qualité par l'analyse visuelle et expérimentale des images des séquences. La métrique subjective est basée sur l'observation de l'être humain. D'après la Figure 26, les régions d'intérêt simplifiées ont une qualité suffisante : la scène est encore compréhensible, une description générale des objets peut être encore tirée et leurs informations importantes sont encore conservées pour une analyse ultérieure. Dans certains cas, lorsque l'objet est éloigné de la caméra ou possède une petite taille (surtout pour les piétons), le filtrage dégrade la qualité de l'objet et ainsi, ses caractéristiques deviennent non suffisamment claires : nous ne pouvons plus obtenir une description physique certaine de l'objet, mais les informations de déplacements sont

conservées. Pour le contexte applicatif visé dans ce travail où les régions d'intérêt sont principalement des véhicules et des piétons, les résultats obtenus sont satisfaisants.

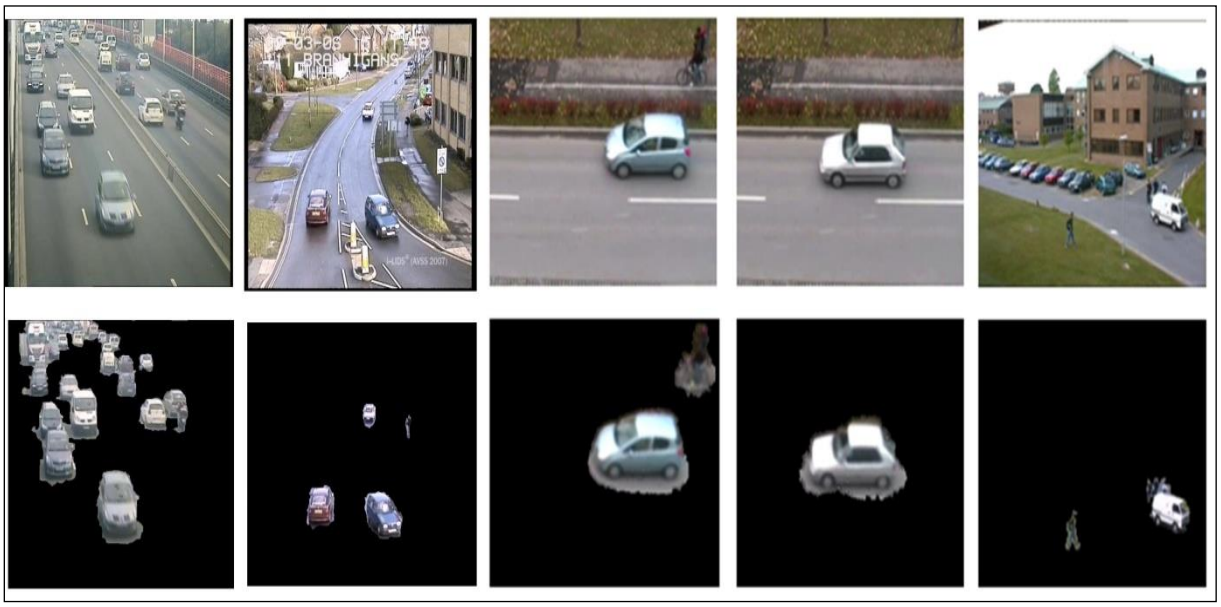

Figure 26. (Haut) Images d'origine. (Bas) Régions d'intérêt spatio-temporellement filtrées

Pour l'évaluation objective, nous utilisons les deux mesures mentionnées le $\boldsymbol{PSNR}$ et la $\boldsymbol{SSIM}$. Les Figures 27, 28 et 29 représentent les valeurs de PSNR calculées pour des séquences originales et simplifiées (Visor, AVSS, video1, Camera1) encodées en utilisant le codeur H.264/AVC à différents débits de compression et paramètres de quantification ($\boldsymbol{QP}$). Comme remarqué précédemment, visuellement, l'image simplifiée a une qualité respectable, mais elle est dégradée par rapport à l'image originale. Les résultats expérimentaux obtenus montrent que les valeurs de $\boldsymbol{PSNR}$, pour la version simplifiée, sont meilleures que l'originale après l'encodage (pour les images $\boldsymbol{I}$ et $\boldsymbol{P}$ séparément dans la Figure 29, pour les séquences entières dans la Figure 28). Particulièrement à des débits d'encodage très faibles (Figure 27 : $\boldsymbol{débit = 5,50\,Kb/s}$ et Figure 28 : $\boldsymbol{QP = 36,48}$), les valeurs de $\boldsymbol{PSNR}$ pour les vidéos simplifiées sont plus importantes que celles d'origine. Pour renforcer les résultats obtenus, la métrique $\boldsymbol{SSIM}$ est calculée. Le degré de similarité entre les images d'origine et simplifiées, avant l'étape de codage, est de $\boldsymbol{70}\%$. Après l'encodage à différents $\boldsymbol{QP}$ (Figure 30.c), les valeurs obtenues de $\boldsymbol{SSIM}$ varient entre $\boldsymbol{64,6}\%$ (à des valeurs élevées de $\boldsymbol{QP}$) et $\boldsymbol{68,1}\%$ (à de faibles valeurs $\boldsymbol{QP}$). La Figure 30 montre également, les résultats de similarité en comparant les images d'origine avant et après codage (Figure 30.a) et les images simplifiés avant et après codage (Figure 30.b). Les valeurs de $\boldsymbol{SSIM}$ diminuent en augmentant la valeur

de $\boldsymbol{QP}$, mais elles deviennent meilleures pour les images simplifiées que celles originales (pour $\boldsymbol{QP} = \boldsymbol{24, 36, 48}$). Il faut noter que des valeurs plus élevées de $\boldsymbol{QP}$ réduisent la qualité visuelle et vice versa. Ainsi, comme déjà prouvé par les résultats de $\boldsymbol{PSNR}$: à des débits de codage faibles, la qualité des séquences simplifiées est meilleure que des originales. Généralement, toutes les valeurs de PSNR et SSIM obtenues sont toujours dans un intervalle respectable qui prouve la qualité acceptable des images filtrées et leur similarité par rapport aux versions originales. Le taux de distorsion, présenté dans la Figure 31, montre que les images filtrées sont moins dégradées au cours du processus de codage que celles non filtrées. Ce constat est logique puisque le filtrage spatio-temporel proposé simplifie les blocs en correspondance de la même manière, ce qui entraîne une baisse des valeurs des erreurs de prédiction et donc une meilleure compression avec moins de dégradation. Les expérimentations objectives appliquées sur les séquences originales et simplifiées après l'encodage H.264/AVC, montrent que visuellement les vidéos filtrées sont meilleures que les originales en particulier à faibles débits de codage (hautes valeurs de quantification). D'ailleurs, coder des séquences filtrées provoque moins de perte en qualité que coder leurs versions originales. En même temps, les caractéristiques des régions d'intérêt sont conservées pour une analyse ultérieure, comme nous montrerons dans les parties suivantes de l'évaluation.

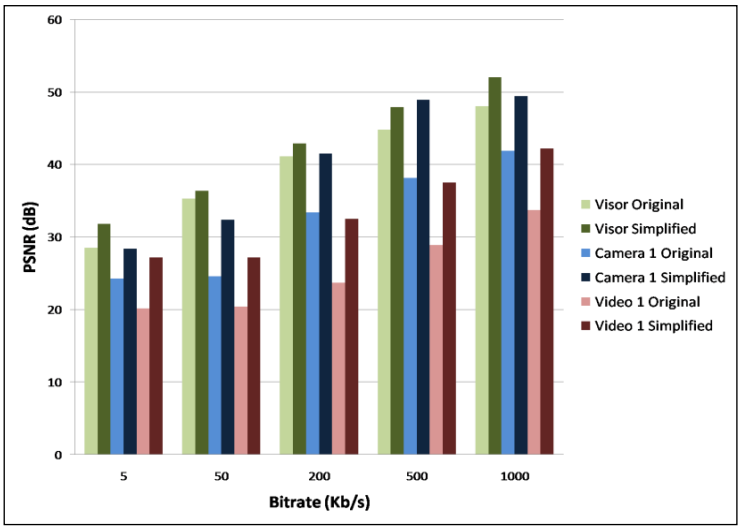

Figure 27. Comparaison des valeurs de PSNR obtenues pour des séquences originales et simplifiées compressées en utilisant l'encodeur H.264/AVC avec un paramètre de quantification fixe QP = 24 et des débits variables de compression

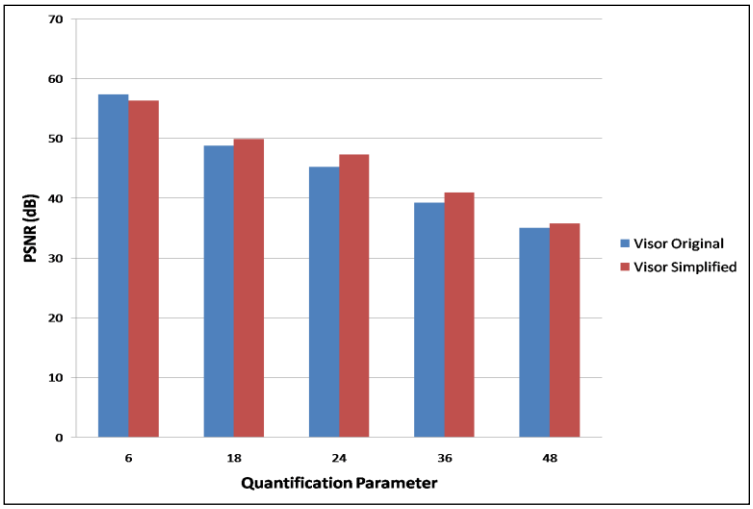

Figure 28. Comparaison des valeurs de PSNR obtenues pour la séquence VISOR originale et simplifiée codée en utilisant l'encodeur H.264/AVC avec des valeurs de paramètres de quantification QP variables

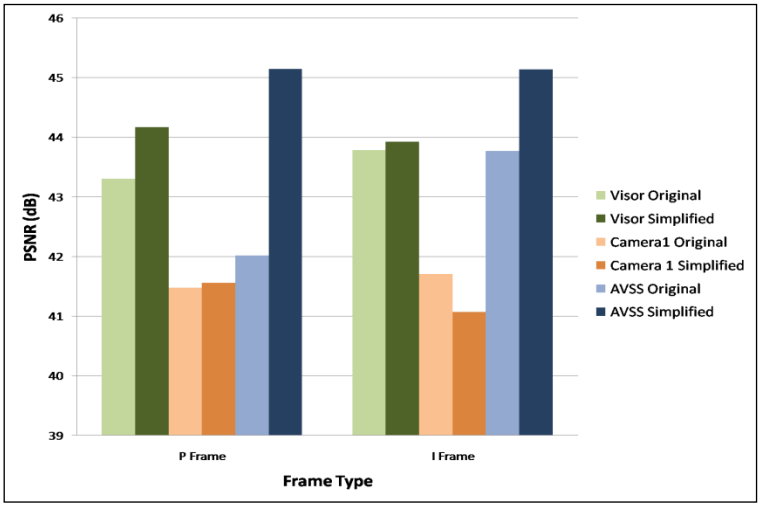

Figure 29. Comparaison des valeurs de PSNR obtenues pour les images I et P des séquences originales et simplifiées compressées en utilisant l'encodeur H.264/AVC avec un paramètre de quantification fixe QP = 24

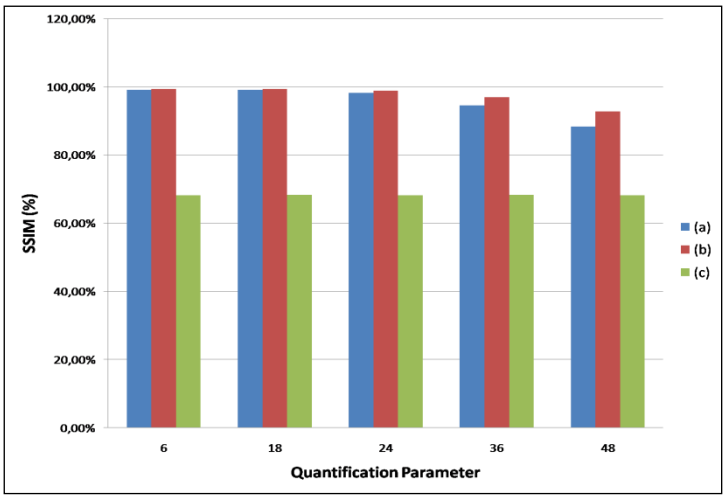

Figure 30. Comparaison des valeurs de SSIM obtenues pour la séquence VISOR originale et simplifiée codée en utilisant l'encodeur H.264/AVC avec des valeurs de QP variables : (a) comparaison entre les images originales et les images originales codées, (b) comparaison entre les images simplifiées et les images simplifiés codées, (c) comparaison entre les images originales et les images simplifiés codées

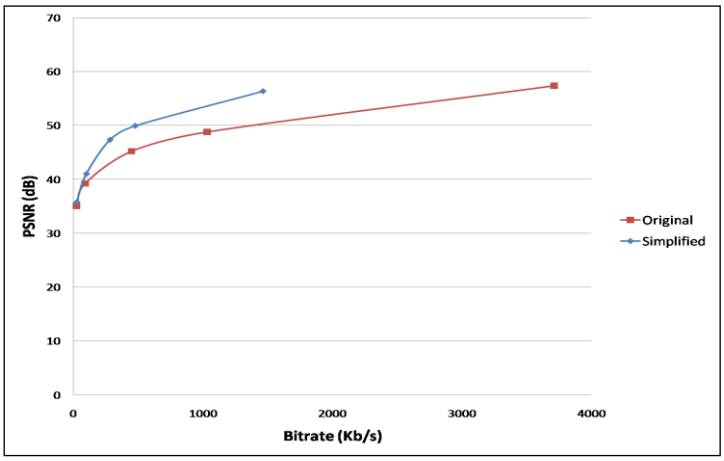

Figure 31. Taux de distorsion de la séquence VISOR originale et simplifiée après l'encodage H.264/AVC

6.5.2. Évaluation computationnelle

Le filtrage spatio-temporel des séquences conduit à un encodage H.264/AVC avec des coûts plus restreints. Dans cette partie, nous comparons les coûts de calcul d'encodage des séquences vidéos originales et simplifiées. Les valeurs des débits obtenues pour la compression de la séquence AVSS à différentes valeurs de paramètres de quantification sont présentées dans la Figure 32, avec les taux de réduction. Sur la Figure 33, les valeurs des débits d'encodage de différentes séquences à $\boldsymbol{QP = 24}$ sont dessinées. Nous pouvons distinguer, à partir de ces deux figures, que les taux de réduction peuvent atteindre plus que $\mathbf{70}\%$ en fonction de la scène et sont meilleurs à faibles valeurs de QP. L'estimation de mouvement est la plus importante et longue étape au cours du processus de codage. Pour cela, nous avons calculé le temps consommé pour l'estimation de mouvement pour des séquences originales et simplifiées. Il est remarquable à partir des Figures 34 et 35 qu'après l'application du filtrage, la consommation en temps de l'estimation mouvement est diminuée avec meilleurs résultats à faibles valeurs de QP. La réduction varie entre $\mathbf{32,15}\%$ et $\mathbf{42,07}\%$ pour des séquences testées à des valeurs moyennes de QP. Le taux de réduction est maximale à un faible QP ($\mathbf{69,17}\%$ à $\boldsymbol{QP = 6}$). Néanmoins à des valeurs très élevées du paramètre de quantification ($\boldsymbol{QP = 48}$), l'étape de filtrage surcharge le processus de codage. Diminuer la consommation en temps de l'estimation du mouvement minimise le temps d'exécution total du processus de codage H.264/AVC. Par conséquent, la compression devient plus rapide avec des coûts de calcul inférieurs. Le temps d'exécution total du processus de codage est calculé dans la Figure 36, et les mêmes conclusions peuvent être extraites avec un taux de réduction

atteignant $\mathbf{48.20}\%$. Pour valider encore l'impact du filtrage spatio-temporel sur l'encodage, nous avons calculé les valeurs des erreurs de prédiction pour chaque macro-bloc lors de l'encodage des séquences dans leurs versions originales et simplifiées. La réduction saillante des erreurs de prédiction après la simplification spatio-temporelle peut être distinguée à partir de la Figure 37. Les valeurs des erreurs de prédiction pour les séquences simplifiées encodées (courbes rouges) sont inférieures par rapport aux valeurs des erreurs de prédiction pour les séquences originales encodées (courbes bleues). Également, elles sont proches de zéro puisqu'elles représentent les erreurs de prédiction pour chaque deux blocs appariés déjà filtrés. En outre, nombreuses valeurs sont nulles parce qu'elles appartiennent à deux macro-blocs exactement similaires. Ceci peut être expliqué par le fait que la phase de filtrage, simplifie chaque deux blocs correspondants de la même manière. Le taux de réduction pour les valeurs des erreurs de prédiction atteint plus que $\mathbf{64,58}\%$ comme le montre la Figure 37. Dans certains cas, comme dans l'exemple illustré dans la Figure 38, les premières valeurs calculées sont très proches de celles originales. En effet, elles sont calculées pour des blocs appartenant à un nouvel objet qui vient d'apparaitre dans la scène (objet encerclé dans l'image jointe). Donc, Ils n'ont pas de blocs correspondant dans l'image précédente. A partir de ces résultats, nous pouvons conclure que la méthode de filtrage spatio-temporel diminue l'erreur de prédiction, ainsi que le débit et le temps de calcul ce qui conduit à une compression meilleure et plus rapide.

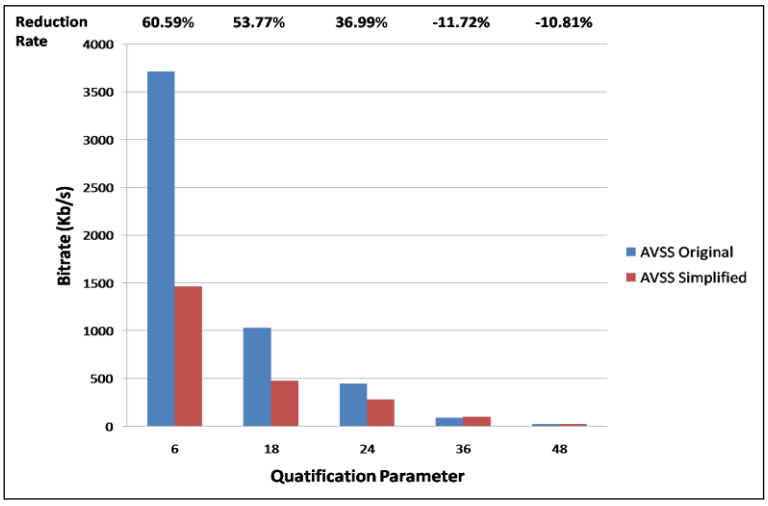

Figure 32. Comparaison des valeurs de débit binaire et des taux de réduction pour la séquence AVSS originale et simplifiée codée à l'aide du codeur H.264/AVC avec différentes valeurs de paramètre de quantification

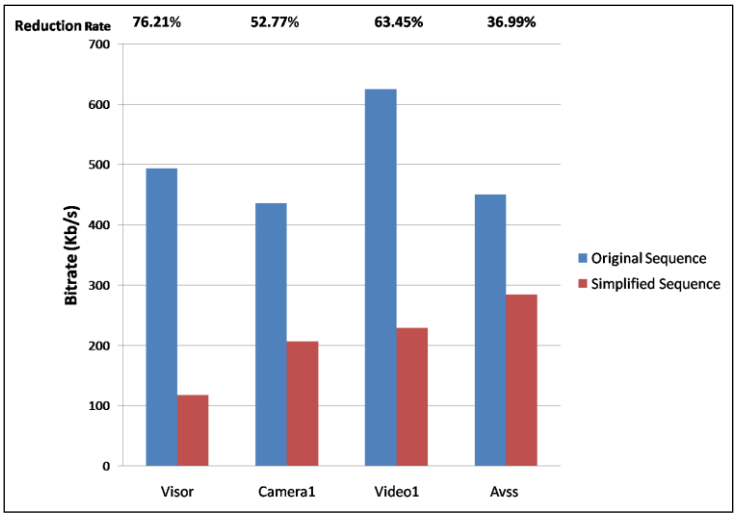

Figure 33. Comparaison des valeurs de débit binaire et des taux de réduction des séquences originales et simplifiées codées à l'aide du codeur H.264/AVC avec QP = 24

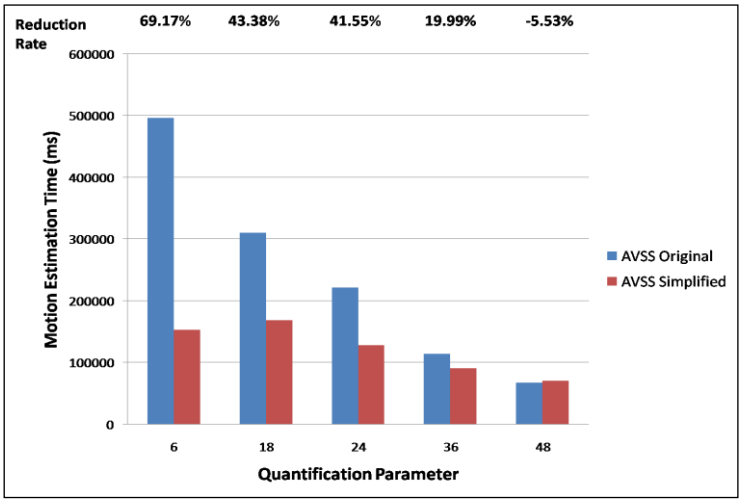

Figure 34. Comparaison du temps d'exécution et des taux de réduction de l'étape d'estimation de mouvement pour la séquence AVSS originale et simplifiée codée à l'aide du codeur H.264/AVC avec différentes valeurs de paramètre de quantification

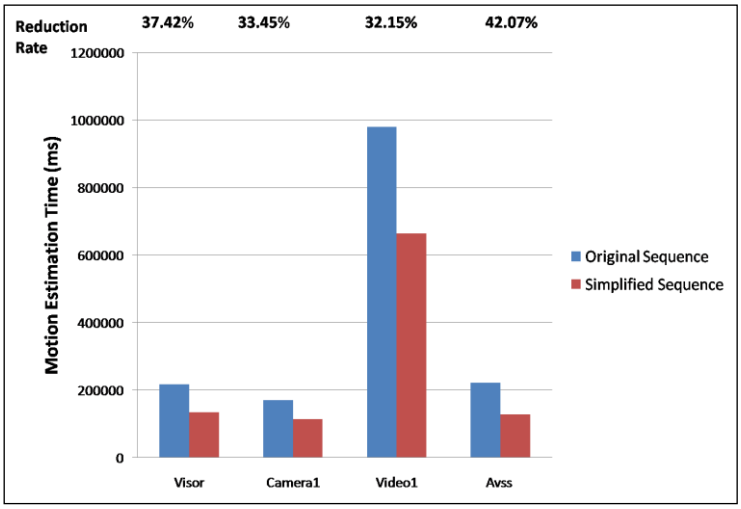

Figure 35. Comparaison du temps d'exécution et des taux de réduction de l'étape d'estimation de mouvement pour des séquences originales et simplifiées à l'aide du codeur H.264/AVC avec QP = 24

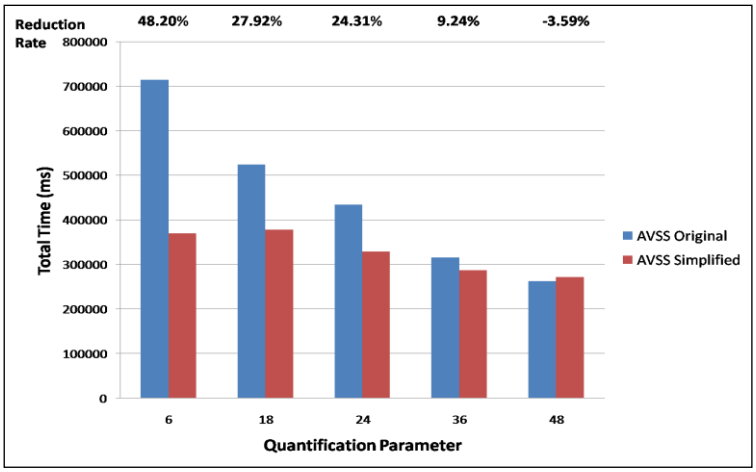

Figure 36. Comparaison du temps d'exécution total et des taux de réduction de l'encodage de la séquence AVSS originale et simplifiée en utilisant l'encodeur H.264/AVC avec QP = 24

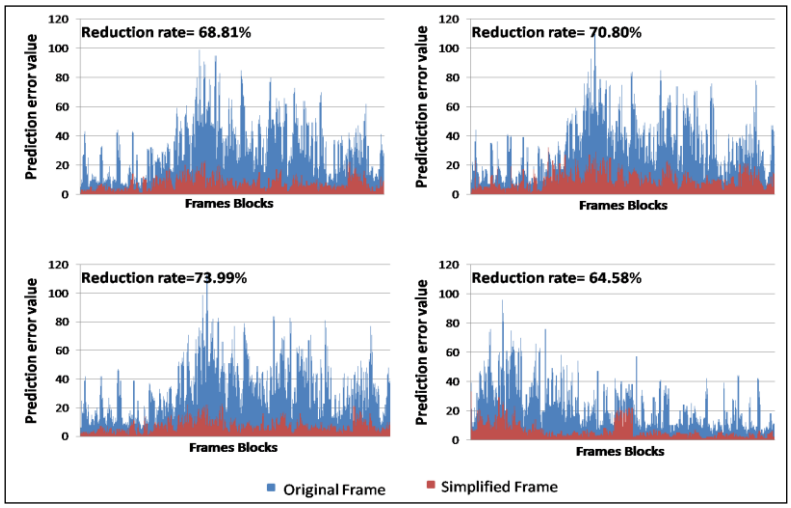

Figure 37. Comparaison des valeurs des erreurs de prédiction calculées pour des images de différentes séquences originales et simplifiées codées en utilisant le codeur H.264/AVC : courbes bleues pour les valeurs des erreurs de prédiction des séquences originales codées, courbes rouges pour les valeurs des erreurs de prédiction des séquences simplifiées codées

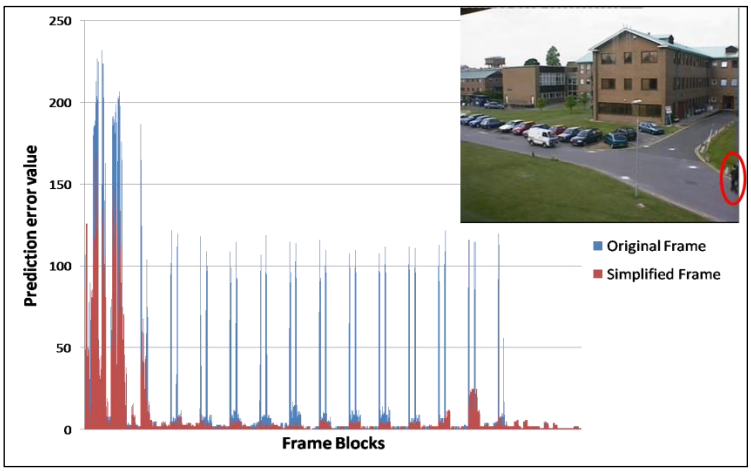

Figure 38. Valeurs des erreurs de prédiction calculées pour des blocs appartenant à un nouvel objet entrant dans la scène pour la séquence Camera1 originale et simplifiée codée en utilisant le codeur H.264/AVC

6.5.3. Évaluation applicative

Pour l'évaluation applicative, nous avons analysé les séquences filtrées spatio-temporellement avec la méthode de suivi d'objets de [Chen 2005] et ensuite, nous avons calculé les valeurs de précision et de rappel. Nous avons comparé les résultats obtenus avec la vérité terrain. Les trajectoires pour chaque objet de la vérité-terrain sont construites à partir du centre des boites englobantes sélectionnées manuellement. Si un objet détecté est associé au même objet de la vérité-terrain pendant la majorité de sa durée de vie, alors il est considéré comme correctement suivi. Si deux objets sont associés au même objet de la vérité-terrain, alors celui ayant le plus grand recouvrement temporel lui est associé et le second objet est considéré comme étant un faux positif.

Différentes séquences originales et simplifiées sont testées avant et après le codage à différentes valeurs de QP. Dans la Figure 39, le rappel pour la séquence simplifiée non encodée ($\boldsymbol{QP = 0}$) est meilleur que sa version originale. Aussi, les valeurs de précision sont plus élevées pour la séquence AVSS simplifiée non encodée. Après le codage H.264/AVC, les valeurs de précision et de rappel à différents QP sont toujours meilleures pour les séquences simplifiées que leurs versions originales. Elles diminuent pour des valeurs de $\boldsymbol{QP}$ dépassant $\boldsymbol{36}$ où les valeurs de débit sont très faibles, de sorte que la qualité des séquences vidéos devient pauvre et la scène n'est plus compréhensible pour suivre les objets en mouvements. Des valeurs de rappel et de précision sont présentes dans la Figure 40 pour le suivi dans diverses séquences codées. Ces valeurs atteignent plus de $\boldsymbol{0,78}$. Elles sont généralement plus élevées pour les versions filtrées. En fait, cela prouve qu'avec le filtrage spatio-temporel, presque tous les objets en mouvement sont correctement suivis sans avoir des faux positifs.

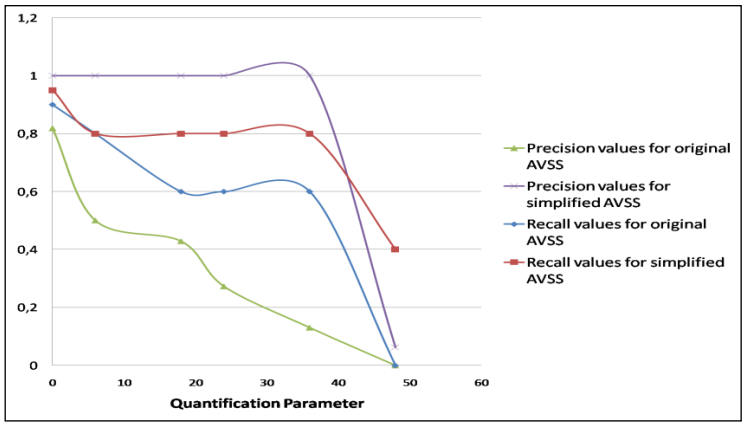

Figure 39. Valeurs de rappel et de précision calculées pour le suivi d'objets en mouvement dans la séquence AVSS originale et simplifiée avant et après codage H.264/AVC avec différentes valeurs des paramètres de quantification

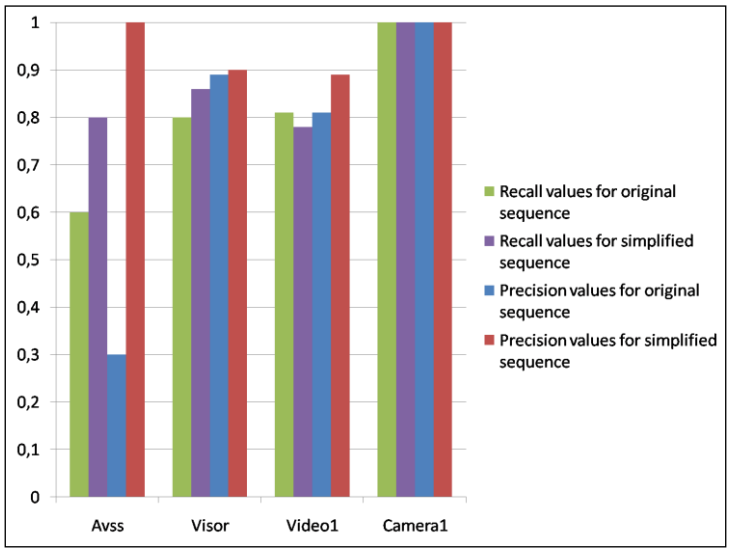

Figure 40. Valeurs de rappel et de précision calculées pour le suivi d'objets en mouvement dans des séquences originales et simplifiées encodées à QP = 24

Le filtrage spatio-temporel décrit ci-dessus est présenté comme un outil de pré-analyse pour les applications de vidéosurveillance. La qualité visuelle de la séquence filtrée est acceptable, tandis que ses redondances spatio-temporelles sont maximisées. Cette méthode élimine les données intelligemment, donc après l'encodage, nous ne maintenons que les plus importantes pour l'analyse. Bien que la méthode proposée soit simple, les résultats montrent que les zones d'intérêt sont simplifiées sans modifier les informations demandées pour une analyse ultérieure. Le processus de codage habituel (sans la méthode de filtrage) élimine les données d'une façon aveugle. Alors qu'avec le filtrage spatio-temporel, nous supprimons intelligemment les informations inutiles en tenant compte des besoins de l'analyse finale. Ainsi, les coûts de l'encodage H.264/AVC sont réduits avec moins de dégradation au niveau de la qualité.

6.6. *Evaluation de la modélisation parallélépipédique d'objets en mouvement*

6.6.1. Évaluation perceptive

La méthode proposée pour la modélisation parallélépipédique peut représenter (ou, au moins, encadrer) différents types d'objets (notamment des véhicules et des piétons) avec le maintien des informations nécessaires pour les applications de la vidéosurveillance. Cette représentation est indépendante de la caméra et de l'objet. Sa simplicité permet aux

utilisateurs de définir facilement un nouvel objet mobile apparaissant dans la scène. Le modèle proposé est défini comme étant un parallélépipède perpendiculaire au plan de l'image. La Figure 41 montre des résultats de modélisation des objets en mouvements. Suivant le contexte applicatif de nos travaux qui est la surveillance routière et des lieux publics, ces objets mobiles sont principalement des véhicules et des piétons. Les modèles obtenus couvrent totalement les objets en mouvement avec un taux de couverture positive de **100**% (pixels appartenant à l'objet et au modèle), ce qui signifie que tous les pixels de l'objet sont englobés par le modèle. Quant au taux de couverture négative, la plus faible valeur (pixels appartenant au modèle et non à l'objet) est inférieure à **20**% pour l'ensemble des objets modélisés. La Figure 42 compare la modélisation des objets par le modèle parallélépipédique à la modélisation par le modèle rectangulaire. Nous pouvons remarquer que le modèle proposé permet de déterminer facilement la classe de l'objet et d'obtenir une bonne approximation des dimensions et de la position de l'objet dans la scène. Aussi, l'orientation du modèle peut être déduite pour définir la direction principale des mouvements de l'objet. En plus, le taux de couverture négative est plus élevé pour le modèle rectangulaire par rapport au modèle parallélépipédique. Dans la Figure 43, le rappel et la précision de la modélisation d'objets par rapport à la vérité terrain (annotation manuelle des objets) sont représentés. Les valeurs obtenues varient entre **0**, **8** et **1** en fonction de la complexité de la scène. Presque tous les objets sont modélisés avec succès, quelques résultats faux positifs sont détectés à cause des occlusions d'objets. D'autres sont des faux négatifs, ils sont ratés à cause de leurs petites tailles.

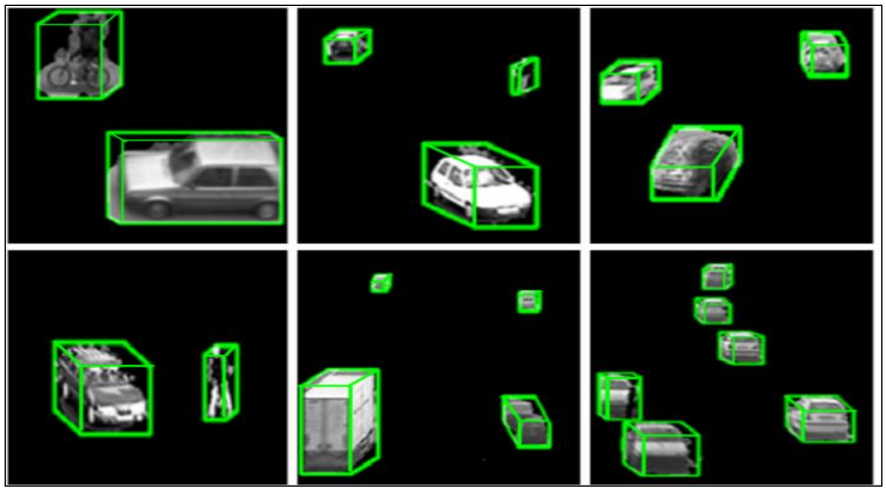

Figure 41. Exemples de résultats de modélisation parallélépipédique d'objets

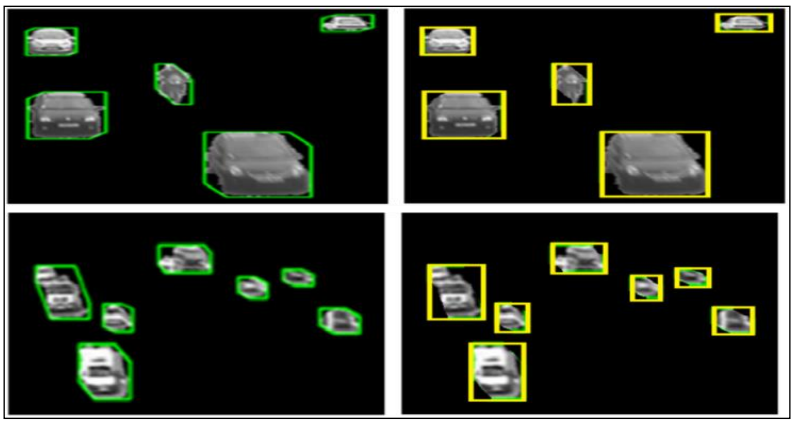

Figure 42. Comparaison entre (Gauche) le modèle parallélépipédique en vert et (Droite) le modèle rectangulaire en jaune

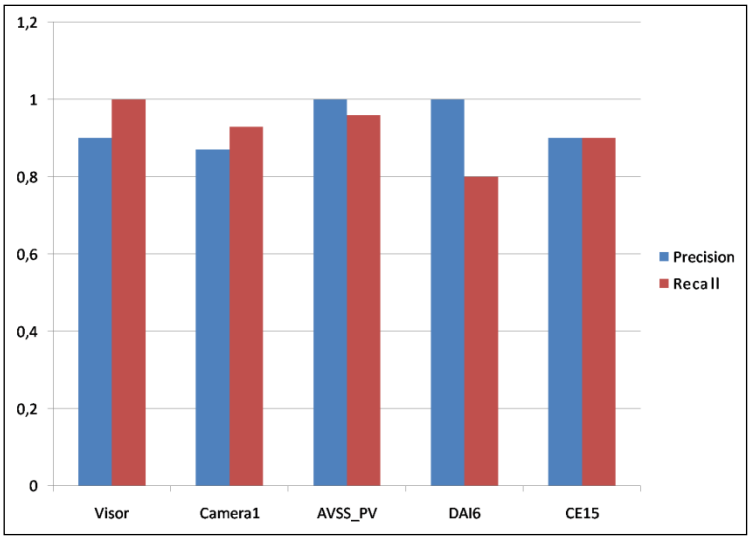

Figure 43. Valeurs de précision et de rappel calculées pour la modélisation parallélépipédique d'objets

6.6.2. Évaluation applicative

Pour prouver que les informations requises pour l'analyse des déplacements de l'objet sont conservées, l'évaluation applicative consiste à suivre les objets modélisés avec la méthode de suivi d'objets de [Chen 2005] et à calculer les valeurs de précision et de rappel. Nous avons comparé les résultats obtenus avec la vérité terrain. Les trajectoires pour chaque objet de la vérité-terrain sont construites à partir du centre des boites englobantes sélectionnées manuellement. Si un modèle détecté est associé au même objet de la vérité-terrain pendant la majorité de sa durée de vie, alors il est considéré comme correctement suivi. Si deux modèles sont associés au même objet de la vérité-terrain, alors celui ayant le plus grand recouvrement temporel lui est associé et le second modèle est considéré comme étant un faux positif.

Différentes séquences en leurs versions originales et modélisées sont testées. Dans la Figure 44, les valeurs de précision et de rappel obtenues atteignent plus de $\mathbf{0,78}$ pour la version modélisée avec une augmentation de $\mathbf{1}\%$ et $\mathbf{11}\%$ par rapport à la version originale. Ceci confirme que presque tous les objets modélisés sont correctement suivis. Par conséquent, la technique de modélisation proposée simplifie les objets, mais en même temps garde les informations requises pour une analyse ultérieure plus détaillée.

En outre, la technique proposée n'impose pas de temps de calcul supplémentaire, elle fonctionne en temps réel. Pour de nombreuses applications de vidéosurveillance, la séquence contenant seulement les modèles d'objets peut être suffisante pour l'objectif de l'application. Dans ce cas, les coûts de traitement et de transmission peuvent être réduits puisque les objets sont remplacés par leurs modèles parallélépipédiques. Ce modèle peut être décrit par les coordonnées de ses six sommets dans le plan de l'image et peut être tout simplement représenté par des métadonnées à partir desquelles des informations sur les dimensions, la position et l'orientation de l'objet peuvent être extraites.

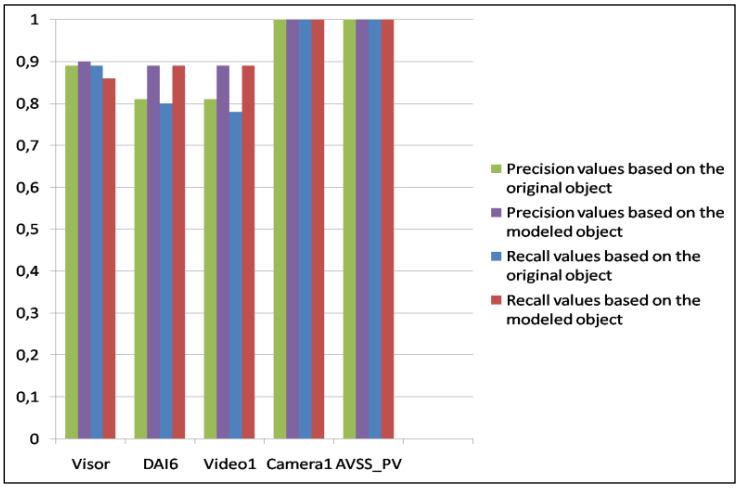

Figure 44. Valeurs de rappel et de précision calculées pour le suivi d'objets en mouvement dans des séquences originales et modélisées

7. Conclusion

Durant ce chapitre, nous avons élargi l'architecture habituelle des systèmes de vidéosurveillance par l'ajout d'une étape de pré-analyse qui extrait les informations pertinentes de la scène. Deux différentes méthodes de pré-analyse ont été proposées. La première méthode consiste en un filtrage spatio-temporel des objets en mouvements qui simplifie les

régions d'intérêt sans altérer les informations demandées pour l'analyse ultérieure. Comme montré expérimentalement, le suivi d'objets est réussi en utilisant la séquence filtrée. En même temps, nous avons accompli une diminution importante des coûts de codage H.264/AVC en réduisant les valeurs des erreurs de prédiction. Quant à la seconde technique, c'est une méthode de modélisation géométrique de la forme qui simplifie la représentation de l'objet en mouvement en un modèle parallélépipédique. Elle permet de maintenir les caractéristiques requises pour le suivi d'objets. Cette représentation est indépendante de la vue de la caméra et de l'orientation de l'objet, elle peut être appliquée sans aucune connaissance préalable des paramètres intrinsèques de la caméra et de l'environnement.

Chapitre 3

Une nouvelle architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance

1. Introduction

Les systèmes de vidéosurveillance ont diverses applications qui peuvent être classées en quatre catégories partant des applications de bas niveau jusqu'à des applications de haut niveau. Conformément à l'application souhaitée par l'utilisateur final, le système exécute une ou plusieurs fonctions analytiques (détection d'objet, suivi d'objet, d'identification et de classification d'objet, l'analyse du comportement). A travers ce chapitre, nous présentons une architecture scalable orientée vers l'application dans le contexte de la vidéosurveillance. Les systèmes proposés sous cette architecture sont multi-niveaux applicatifs, contrairement à ceux habituels qui sont à un seul niveau d'application. Parallèlement à la scalabilité du codage des flux vidéo, le concept de la scalabilité est intégré dans cette approche pour que le niveau applicatif du système puisse être incrémenté progressivement suivant les besoins de l'utilisateur final. Le processus de cette architecture se déroule de façon scalable pour offrir la possibilité aux systèmes de répondre à plusieurs applications de différents niveaux.

Dans la première section de ce chapitre, nous décrivons l'approche de la scalabilité orientée vers l'application pour les systèmes de vidéosurveillance. Également, nous proposons un exemple de système qui respecte cette architecture où les deux méthodes de pré-analyse déjà présentées (dans le Chapitre 2) sont intégrées. Quant à la dernière section (Section 3), elle évalue les performances de cette approche et étudie les effets de la combinaison de la scalabilité orientée vers l'application avec les dimensions temporelles et spatiales du codage de flux vidéo scalable.

2. Architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance

Les contributions déjà développées dans les chapitres précédents sont exploitées dans cette architecture en l'adaptant avec le concept de la scalabilité. Une présentation des travaux connexes sur le rôle de la scalabilité dans la vidéosurveillance et leurs limitations est fournie pour justifier les motivations de cette idée. Par la suite, une description de l'approche est détaillée.

2.1. *Travaux connexes de scalabilité dans les systèmes de vidéosurveillance*

La scalabilité dans la vidéosurveillance a été toujours étudiée comme un outil pour renforcer les capacités de codage, de transmission et de traitement des flux vidéo dans les

systèmes de vidéosurveillance. Le codage vidéo scalable a prouvé son efficacité pour faciliter différentes fonctions en vidéosurveillance telles que : le suivi à distance en temps réel, la politique de stockage, l'analyse de l'image, la surveillance interactive,... [Ziliani 2005]. Plusieurs travaux, dans la littérature, utilise le codage vidéo scalable pour des objectifs de surveillance comme : Lambert et al., dans [Lambert 2006], l'utilisent pour le streaming en temps réel; Wang et al., dans [Shizheng 2011], proposent le codage vidéo scalable comme un outil efficace pour stocker et chercher les synopsis vidéo; les auteurs de [Grois 2010] l'exploitent pour respecter les limites de la bande passante lors de la transmission des flux vidéo. Le codage vidéo scalable est utile aussi pour la sécurité et la protection : Il peut fournir une surveillance sécurisée, comme indiqué dans le travail de Parc [Wan 2011]; une authentification pour assurer l'intégrité des flux vidéo de surveillance [Zhuo 2013]; la protection de la vie privée des personnes dans les vidéos transmises [Hosik 2009]. En la combinant avec d'autres approches telles que le cloud computing [Feng 2012] ou les réseaux de capteurs [Yongil 2010], la scalabilité améliore les performances des systèmes de vidéosurveillance. Dans le travail de [Detmold 2006], le concept de la scalabilité est appliqué dans la construction des réseaux de la vidéosurveillance à grande échelle pour assurer un stockage et une transmission scalables et distribués. Également, les auteurs de [Albusac 2011] propose une approche de vidéosurveillance scalable qui utilise des composantes de normalité pour augmenter les capacités d'analyse. À travers la littérature, la relation entre la scalabilité et la vidéosurveillance a été généralement exploitée pour le traitement des données vidéo indépendamment de l'application finale du système de vidéosurveillance.

2.2. *Schéma de l'approche proposée*

Dans ce travail, nous avons pris un autre point de vue envers cette relation. Nous avons étudié la scalabilité dans le contexte de l'application et nous proposons, alors, une architecture de vidéosurveillance scalable orientée vers l'application. La scalabilité, ici, est représentée dans la hiérarchie du système qui dépend de l'application finale demandée. Cette dernière est choisie par l'utilisateur final et ainsi, le résultat désiré de l'application influence sur la phase d'analyse en identifiant les tâches à réaliser. L'analyse de la scène enregistrée peut être effectuée à différents niveaux d'abstraction partant de la fonction de bas niveau (la détection d'objet), jusqu'à la fonction de haut niveau (compréhension du comportement). Par

conséquent, les informations pertinentes à extraire et à traiter varient également selon l'application.

Dans le codage de vidéo scalable, une faible fréquence d'images ou une basse résolution du flux vidéo est d'abord traitée pour former la couche de base de la vidéo codée. Une deuxième couche d'information, appelée couche de rehaussement, est ensuite construite à partir d'une fréquence temporelle ou une résolution supérieure du flux vidéo en utilisant la couche de base pour guider le processus de codage. La troisième couche est codée de la même manière en utilisant la seconde couche en tant que référence. Ce processus se poursuit sur toutes les couches successives. L'avantage de cette approche est que le dispositif du client peut décoder le flux reçu, en commençant par la couche de base, puis décoder les informations supplémentaires à partir des couches de rehaussement successives jusqu'à ce que la fréquence temporelle ou la résolution souhaitée soit atteinte. Un appareil, ayant un écran de faible résolution ou des puissances de calcul limitées pour le décodage, peut choisir d'arrêter le décodage après les premières couches. Un autre dispositif de haute définition ou avec des capacités élevées peut décoder toutes les couches obtenant ainsi la vidéo est en pleines résolution et fréquence temporelle. De cette façon, un seul flux peut être utilisé pour servir tout type d'appareils en permettant de décider le nombre de couches à décoder.

De façon similaire au processus de codage vidéo scalable, notre idée principale est une scalabilité progressive selon l'application souhaitée par l'utilisateur. Le concept de la scalabilité est appliqué sur l'architecture du système de vidéosurveillance. La couche de base contient des informations pour les applications de surveillance de bas niveau. Le second flux, qui est une couche de rehaussement, emporte des données pour une application de niveau supérieur. Les flux de rehaussement subséquents fournissent des informations pour des niveaux plus élevés. À chaque fois, les données nécessaires sont extraites, codées pour construire la couche correspondante et ensuite transmises. L'utilisateur final reçoit en premier temps la couche de base. Après son décodage, l'analyse finale interprète ce flux pour générer les résultats attendus de l'application. Si le client est à la recherche de résultats plus détaillés, il doit atteindre une application de niveau supérieur. Ainsi, une requête est envoyée et le flux de rehaussement correspondant est transmis. En le combinant avec les couches précédemment envoyées, une application plus avancée est atteinte et ,donc, des résultats plus élaborés sont obtenus. Ce processus peut être appliqué de manière récursive pour améliorer le niveau de l'application jusqu'à ce que la sortie atteinte réponde aux besoins de l'utilisateur final.

Citons par exemple, les systèmes de surveillance de stationnement dans les parkings qui visent généralement à détecter/compter les véhicules entrant et sortant, les suivre afin de

s'assurer qu'ils sont bien garés et alerter, sinon. Distinctement, la phase d'analyse de ces systèmes est basée sur les deux fonctions de bas niveau : la détection et le suivi d'objets. En cas de stationnement dans une zone interdite, le système alerte automatiquement l'utilisateur. Dans certain cas, il est nécessaire d'identifier le véhicule alors qu'il est irréalisable avec ce type de systèmes. Les informations extraites pour le suivi d'objets ne sont plus suffisantes et ainsi, révéler l'identité du véhicule exige une application plus avancée. Avec l'architecture que nous proposons, si l'utilisateur demande d'avoir les détails de l'objet (type, modèle, plaque d'immatriculation), les informations supplémentaires sont récupérées et transférées afin de réaliser l'identification de l'objet.

L'architecture scalable suggérée est décrite dans la Figure 45. Les informations nécessaires pour la couche de base $(\boldsymbol{I_0})$ sont extraites, codées, transmises ensuite décodées $(\check{\mathrm{I}}_0)$ à la réception pour être analysé par la fonction de plus faible niveau. Une fois le niveau suivant est demandé, les données nécessaires $(\boldsymbol{I_N})$ sont préparées : la différence $(\boldsymbol{E_N})$ entre les informations précédentes et demandées forme la couche de rehaussement qui va être codée et transmise. Après son décodage, $(\check{\mathrm{E}}_N)$ est ajoutée à $(\check{\mathrm{I}}_{N-1})$ pour former les caractéristiques nécessaires $(\check{\mathrm{I}}_N)$ pour cette application. Cette technique d'envoyer seulement la différence d'informations et d'utiliser les informations précédemment envoyées pour guider l'analyse subséquente réduit les coûts du système. Contrairement aux systèmes de surveillance habituels qui gèrent tous les flux vidéo enregistrés (où de nombreuses régions ne sont pas vraiment intéressantes) pour atteindre une seule application précise, développer des systèmes de surveillance respectant cette architecture proposée évite de surcharger les systèmes de vidéosurveillance. En fait, les informations pertinentes sont progressivement extraites, envoyées et traitées suivant les exigences ascendantes de l'utilisateur final.

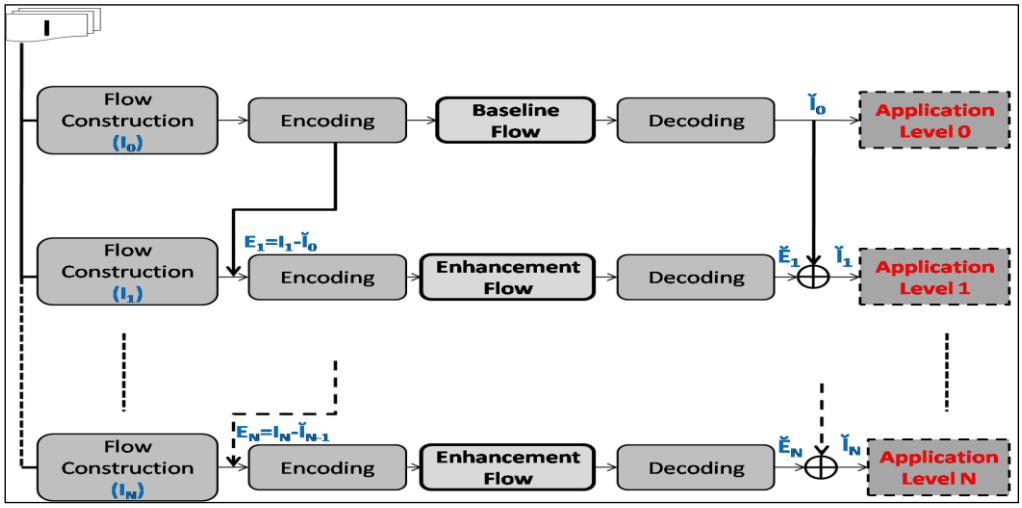

Figure 45. Architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance

2.3. *Schéma du système scalable orienté vers l'application proposé*

En se basant sur l'architecture scalable orientée vers l'application décrite, un exemple de système de vidéosurveillance est proposé. Ce système permet de réaliser les quatre fonctions d'analyse (citées dans la Section 2.2.4 du Chapitre 1) suivant le niveau de l'application demandée. Il peut détecter un objet en mouvement dans la scène comme il peut analyser son comportement. La couche de base contient des informations pour la fonction de détection d'objet. En plus, trois couches de rehaussement sont construites où chacune d'entre elles contient les données nécessaires pour une fonction d'analyse précise. Pour chaque niveau d'application, une étape de pré-analyse extrait les données pertinentes pour les tâches à exécuter. Ainsi, dans le premier niveau, chaque objet en mouvement est extrait puis substitué par son bloc central rempli de la couleur dominante de l'objet (voir Figure 46). De cette façon, le comptage d'objets peut être réalisé en comptant les blocs passants dans la scène. Chaque objet est caractérisé par les positions des quatre sommets de son bloc central et de l'information couleur. Par conséquent, la couche de base est formée seulement de métadonnées à transmettre et aucune technique de codage vidéo n'est nécessaire. La première couche de rehaussement est conçue pour le second niveau applicatif : elle emporte les informations de suivi d'objets. Chaque objet en mouvement est représenté par son modèle parallélépipédique développé par la méthode de modélisation décrite précédemment (Section 5 du Chapitre 2). Les caractéristiques requises pour le suivi d'objets sont conservées (voir Figure 46). Le modèle de chaque objet peut être décrit par les positions des six sommets (la projection du modèle parallélépipède sur le plan de l'image). Donc, également cette couche est constituée de métadonnées et le codage vidéo y est inutile. Quant à l'application du troisième niveau, les informations demandées sont transmises à travers la seconde couche de rehaussement pour assurer la tâche de classification d'objets. Dans ce niveau, nous utilisons la méthode de filtrage spatio-temporel déjà présentée (Section 4 du Chapitre 2) pour simplifier les objets en mouvement. Ainsi, les informations nécessaires pour la classification sont maintenues (voir Figure 46). Le flux de rehaussement est formé par les objets simplifiés codés. L'encodage H.264/AVC est appliqué sur les régions simplifiées ce qui conduit à une meilleure compression avec des coûts réduits. L'application de plus haut niveau, doit permettre d'identifier les comportements des objets d'intérêt. Pour ce niveau, les régions d'intérêt sont maintenues dans leur format d'origine sans être simplifiées puisque la tâche d'analyse des comportements a besoin de plusieurs caractéristiques pour réussir l'interprétation. Ce flux de rehaussement contient les objets en mouvement dans leur qualité

d'origine et, ainsi, le codage H.264/AVC est nécessaire. Le système fonctionne d'une manière scalable : Tout d'abord, la couche de base est envoyée pour le comptage d'objets. Ensuite, chaque fois que l'utilisateur demande d'atteindre un niveau supérieur, la couche de rehaussement correspondante est envoyée. Le flux ne contient que la différence entre les données des deux niveaux successifs ce qui assure des coûts de codage et de transmission réduits. Dans les systèmes courants de vidéosurveillance, les données traitées sont toujours les images complètes des scènes filmées. Tandis que dans notre système proposé, les données sont réduites aux caractéristiques requises comme expliqué dans le Tableau 1. Grâce à la phase de pré-analyse, les régions d'intérêt sont extraites puis simplifiées pour ne conserver que les informations nécessaires en fonction de la tâche sélectionnée.

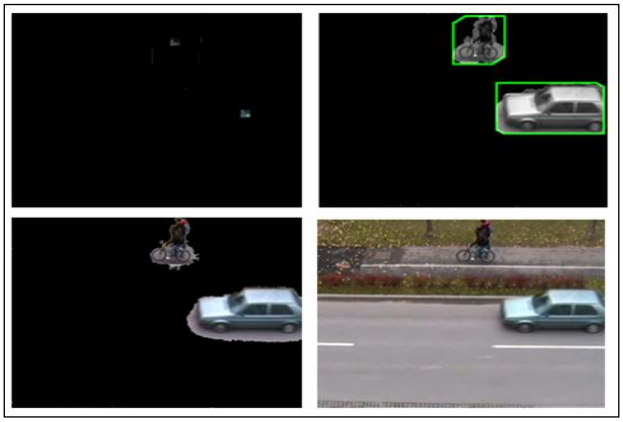

Figure 46. (Haut gauche) Informations extraites pour le premier niveau : blocs centraux des objets. (Haut droite) Informations extraites pour le deuxième niveau : modèles parallélépipédiques des objets. (Bas gauche) Informations extraites pour le troisième niveau : les objets en mouvement simplifiés spatio-temporellement. (Bas droite) Informations extraites pour le quatrième niveau : séquence dans sa qualité originale

Tableau 1. Les quatre niveaux du système proposé sous l'architecture scalable orientée vers l'application

Niveau applicatif	**Type de la couche**	**Caractéristiques analysées**	**Type des caractéristiques**	**Tâche analytique**	**Application**
1	Base	Positions des sommets du bloc central de l'objet + information couleur dominante	Métadonnées	Détection d'objets	Comptage des objets en mouvements
2	Rehaussement	Positions du modèle de l'objet	Métadonnées	Suivi d'objets	Suivi des objets en mouvements
3	Rehaussement	Objet simplifié	Données encodées	Classification d'objets	Distinction entre véhicules et piétons
4	Rehaussement	Objet brut	Données encodées	Analyse des comportements	Détection d'événements anormaux

3. Résultats expérimentaux

Les expérimentations sont réalisées dans les mêmes conditions et avec les mêmes outils que les expérimentations du chapitre précédent. Le système proposé dans la Section 2.3 est utilisé pour évaluer la scalabilité orientée vers l'application dans le contexte de la vidéosurveillance. Il est comparé avec un système de vidéosurveillance habituel qui est le système proposé par la bibliothèque OpenCV [Chen 2005]. Comme expliqué précédemment, les deux dernières couches du système scalable nécessitent une étape d'encodage. De ce fait, la scalabilité spatiale et temporelle du codage vidéo est appliquée sur ces niveaux afin de juger son influence sur la scalabilité proposée. Les évaluations perceptuelle, computationnelle et applicative sont présentées dans les parties suivantes.

3.1. *Évaluation de l'architecture scalable orientée vers l'application*

Pour l'évaluation applicative, la précision et le rappel sont calculés pour chaque niveau applicatif. Pour le premier niveau, la Figure 47 montre les valeurs de précision et de rappel pour le comptage d'objets dans différentes séquences. Nous comparons les résultats avec la vérité terrain, qui consiste en une détection et comptage manuels des objets de la scène. Les valeurs du comptage basé sur le bloc central de l'objet sont presque égales aux valeurs obtenues pour le comptage basé sur l'objet tout entier. Pour l'application de suivi d'objet du deuxième niveau, les valeurs obtenues dans la Figure 48 prouvent que la plupart des régions en mouvement sont correctement suivies avec de meilleurs résultats pour la version modélisée par rapport à l'originale (la vérité terrain est obtenue de la même façon expliquée dans le Chapitre 2). Le troisième niveau applicatif classifie les objets en piétons et véhicules. La vérité terrain consiste en une classification manuelle des objets de la scène . Dans la Figure 49, les résultats de précision et de rappel pour la classification des objets spatio-temporellement filtrés sont égaux ou parfois supérieurs à la classification des objets non simplifiés. Aucune comparaison n'est adressée pour l'analyse des comportements parce que les objets sont conservés sans changement. Depuis que la phase de pré-analyse réduit la quantité d'informations à traiter sans altérer les caractéristiques pertinentes, les tâches analytiques du système testé présentent de bons résultats d'analyse avec des coûts de codage, de transmission et d'analyse plus faibles par rapport aux systèmes habituels.

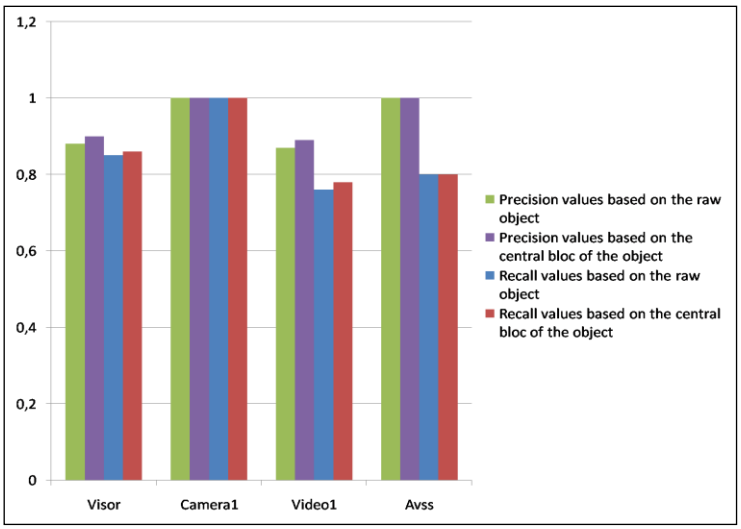

Figure 47. Valeurs de rappel et de précision calculées pour le comptage d'objets en mouvement basé sur le bloc central de l'objet et l'objet entier pour différentes séquences

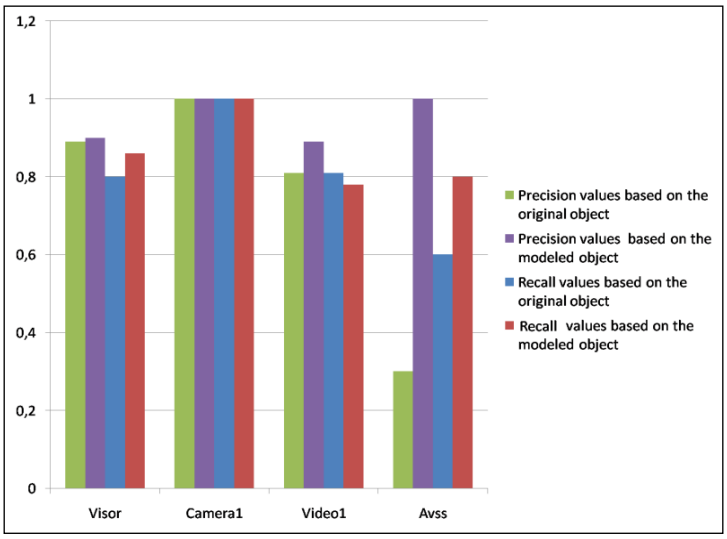

Figure 48. Valeurs de rappel et de précision calculées pour le suivi d'objets en mouvement basé sur le modèle parallélépipédique de l'objet et l'objet original pour différentes séquences

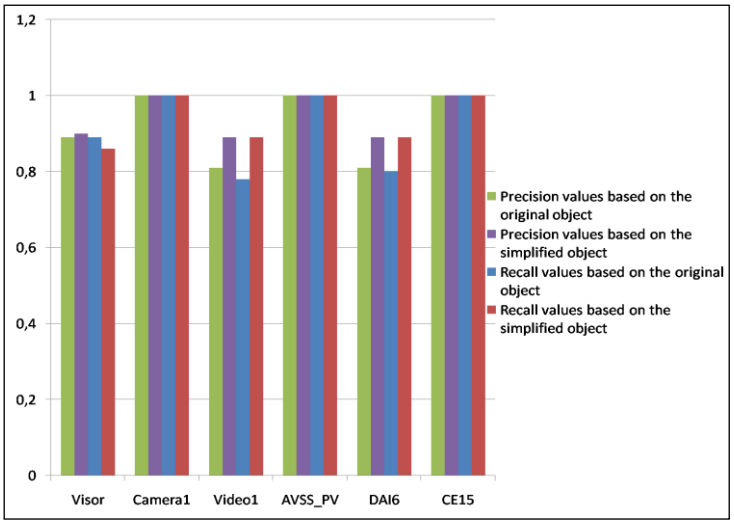

Figure 49. Valeurs de rappel et de précision calculées pour la classification d'objets en mouvement basé sur l'objet spatio-temporellement filtré et l'objet original pour différentes séquences

3.2. *Combinaison de la scalabilité orientée vers l'application avec la scalabilité spatiale*

Le format initial des séquences vidéos utilisées est le format $\boldsymbol{CIF}$ ($\boldsymbol{352x288}$). Les expérimentations sont réalisées en variant la résolution spatiale des images pendant la phase d'encodage. Nous comparons entre les séquences traitées par le système de vidéosurveillance ordinaire (où la séquence est entièrement codée, envoyée et analysée) et le troisième niveau du système scalable proposé. Pour évaluer la qualité visuelle, les valeurs de $\boldsymbol{PSNR}$, obtenues dans la Figure 50, sont meilleures pour le système proposé. En particulier, elles sont beaucoup plus élevées, en réduisant la résolution spatiale, grâce à l'utilisation du filtrage spatio-temporel qui garantit une faible dégradation de la qualité des images.

Le débit binaire est également évalué à différentes résolutions dans la Figure 51. Nous pouvons distinguer la réduction du débit dans notre système par rapport au système habituel. Ceci est attendu puisque dans notre approche, nous traitons uniquement les données nécessaires en augmentant leurs redondances spatio-temporelles. Les valeurs du débit binaire sont supérieures pour l'agrandissement de la résolution spatiale avec un taux de réduction de $\boldsymbol{55,78}\%$ pour le format $\boldsymbol{9CIF}$. Pour une très petite résolution ($\boldsymbol{SQCIF}$), le débit de codage du système habituel est meilleur que le nôtre. Pendant que pour le format $\boldsymbol{QCIF}$, qui est largement utilisé, la réduction est assurée avec un taux de $\boldsymbol{8.29}\%$. Le temps d'exécution total est présenté dans la Figure 52. Le meilleur taux de réduction est $\boldsymbol{16,27}\%$ pour le format $\boldsymbol{QCIF}$. La moyenne de bits par image est aussi calculée pour prouver à quel point l'information est réduite dans le système scalable proposé. Comme montré dans la Figure 53, les taux de réduction varient de $\boldsymbol{37,65}\%$ à $\boldsymbol{55,54}\%$ selon la résolution. Les erreurs de prédiction au cours de l'étape d'encodage à différentes résolutions sont également calculées pour la séquence VISOR, dans la Figure 54. Nous pouvons distinguer que la réduction est toujours présente avec de bons taux obtenus.

Pour l'évaluation applicative, nous analysons des séquences à différentes résolutions. Mais comme le montre la Figure 55, la scalabilité spatiale n'a aucun effet sur l'analyse finale : les mêmes valeurs de précision et de rappel sont obtenues pour tous les formats spatiaux. En même temps, elles sont plus élevées avec notre système proposé. À partir de ces résultats, nous pouvons remarquer que la combinaison de la scalabilité spatiale avec l'approche proposée n'a aucun effet négatif sur les aspects perceptuel, computationnel et applicatif du système. Les gains obtenus à partir de la scalabilité orientée vers l'application sont conservés. Ainsi, l'approche développée s'adapte bien avec la scalabilité spatiale.

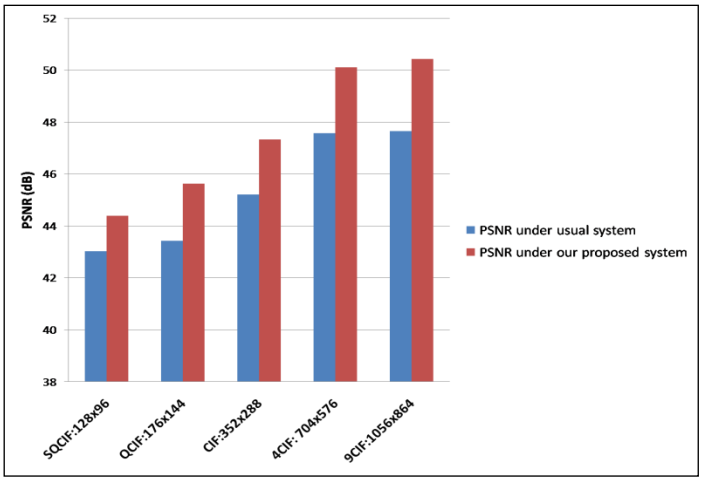

Figure 50. Comparaison des valeurs obtenues de PSNR pour la séquence VISOR à différentes résolutions avec un système de surveillance habituel et le système scalable orienté vers l'application

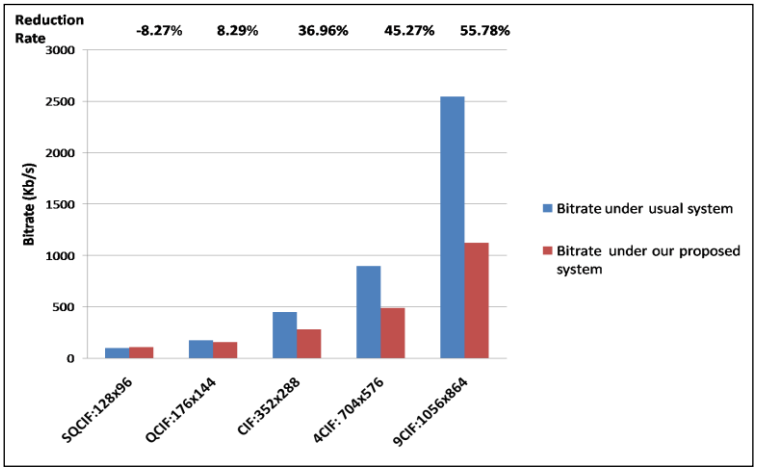

Figure 51. Comparaison des valeurs de débit binaire obtenues pour la séquence VISOR à différentes résolutions avec un système de surveillance habituel et le système scalable orienté vers l'application

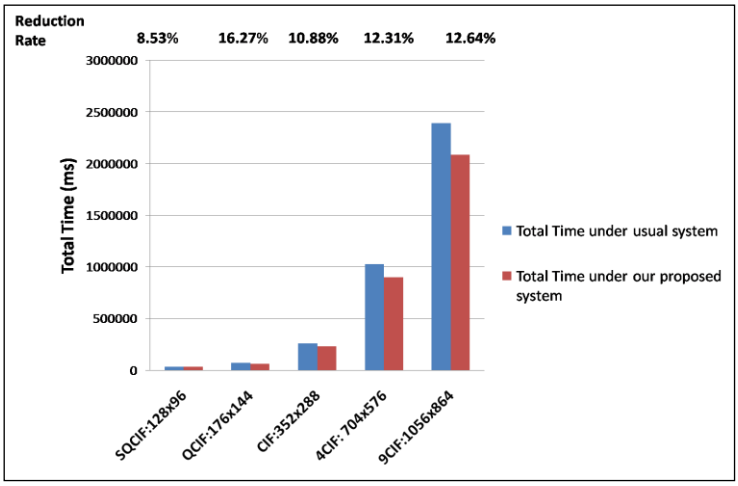

Figure 52. Comparaison des valeurs obtenues du temps d'exécution total pour la séquence VISOR à différentes résolutions avec un système de surveillance habituel et le système scalable orienté vers l'application

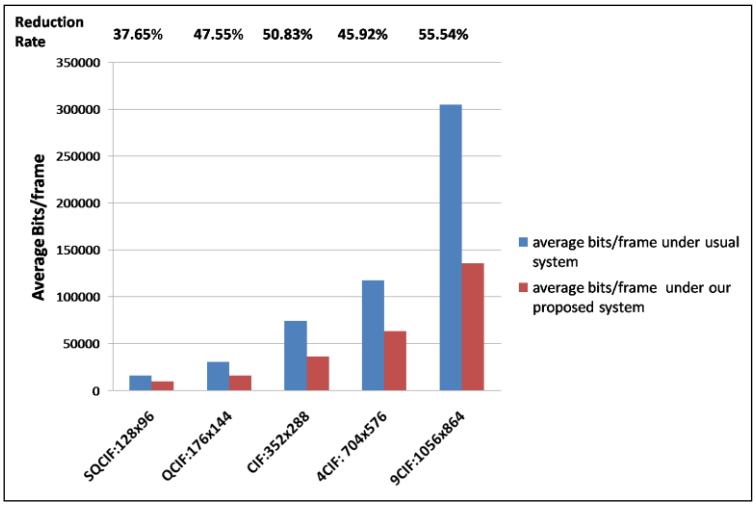

Figure 53. Comparaison des valeurs obtenues de la moyenne de bits par image pour la séquence VISOR à différentes résolutions avec un système de surveillance habituel et le système scalable orienté vers l'application

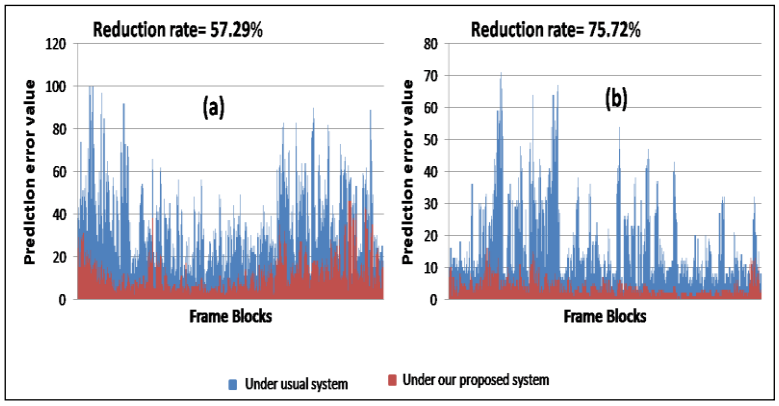

Figure 54. Comparaison des valeurs d'erreur de prédiction calculées pendant l'encodage de la séquence VISOR au (a) format QCIF (176x144), (b) format 4CIF (704x576) avec un système de surveillance habituel et le système scalable orienté vers l'application

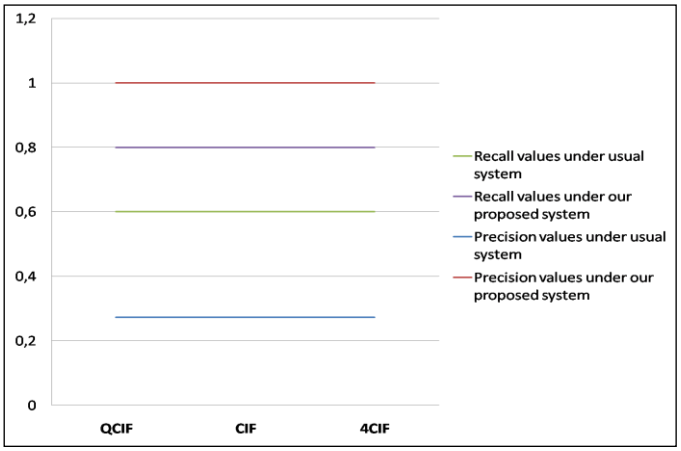

Figure 55. Comparaison des valeurs de précision et de rappel pour l'analyse de la séquence VISOR à différentes résolutions avec un système de surveillance habituel et le système scalable orienté vers l'application

3.3. *Combinaison de la scalabilité orientée vers l'application avec la scalabilité temporelle*

D'une façon similaire à la scalabilité spatiale, l'impact de la scalabilité temporelle sur l'approche proposée est testé. Les expérimentations sont réalisées en changeant la fréquence

temporelle des séquences lors de la phase d'encodage. Il faut noter que la scalabilité temporelle est fixée par le nombre d'images à ignorer (par exemple pour la valeur **2**, nous encodons chaque troisième image). Pour la qualité visuelle, la Figure 56 montre que les valeurs de $\boldsymbol{PSNR}$ de la séquence traitée avec notre système sont meilleures même après un grand nombre d'images. Les valeurs du débit binaire sont également améliorées avec notre approche, comme le montre la Figure 57. Les taux de réduction varient de $\mathbf{48,07}\%$ (nombre d'images à ignorer $=$ **1**) à $\mathbf{61.01}\%$ (nombre d'images à ignorer $=$ **14**). Le temps d'exécution total est également réduit dans la Figure 58. Les taux de réduction sont compris entre $\mathbf{35,07}\%$ et $\mathbf{46,05}\%$. Les valeurs des erreurs de prédiction et de la moyenne de bits par images sont aussi évaluées. Les mêmes conclusions que la scalabilité spatiale peuvent être tirées : Même en le combinant avec la scalabilité temporelle, les performances du système proposé sont meilleures que celui habituel. L'étape d'analyse est réalisée comme le prouve les valeurs de précision et de rappel dans la Figure 59. La précision du système développé est meilleure. Mais pour les deux systèmes, elle devient nulle pour un nombre d'images à ignorer dépassant **9** images. Également, les valeurs de rappel pour l'analyse réalisée par notre système sont plus élevées en passant une seule image (en codant une image toutes les deux). Les valeurs suivantes sont égales jusqu'à s'annuler pour un nombre d'images à ignorer dépassant **9** images. Cela peut être justifié expérimentalement par le fait que la compréhension (et donc l'analyse) de la séquence est difficile au-delà d'une valeur de **2** images à passer et devient impossible à partir de **9** images (les mouvements d'objets deviennent très rapides).

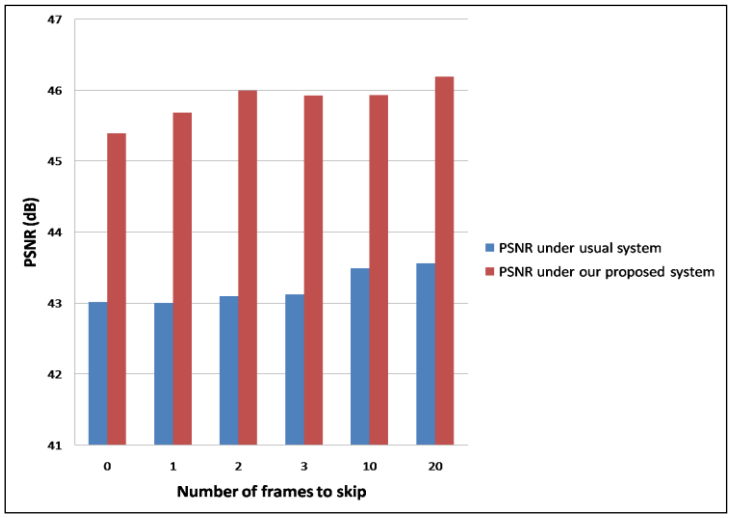

Figure 56. Comparaison des valeurs obtenues de PSNR pour la séquence VISOR en variant le nombre d'images à ignorer avec un système de surveillance habituel et le système scalable orienté vers l'application

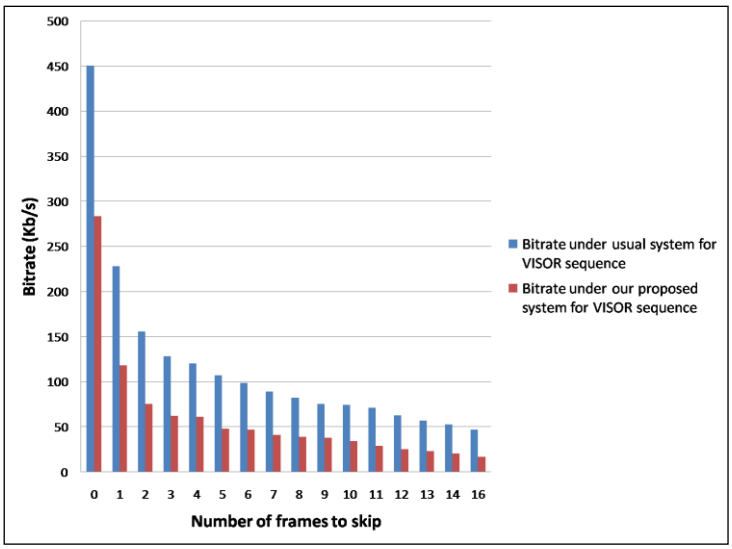

Figure 57. Comparaison des valeurs obtenues de débit binaire pour la séquence VISOR en variant le nombre d'images à ignorer avec un système de surveillance habituel et le système scalable orienté vers l'application

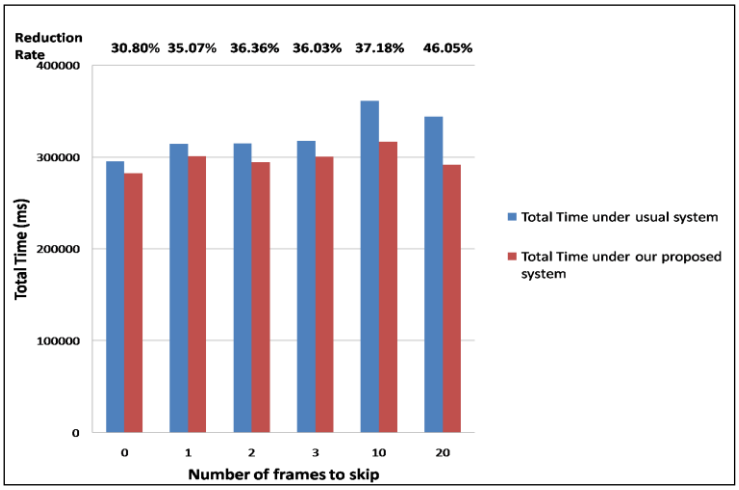

Figure 58. Comparaison des valeurs obtenues de temps d'exécution pour la séquence VISOR en variant le nombre d'images à ignorer avec un système de surveillance habituel et le système scalable orienté vers l'application

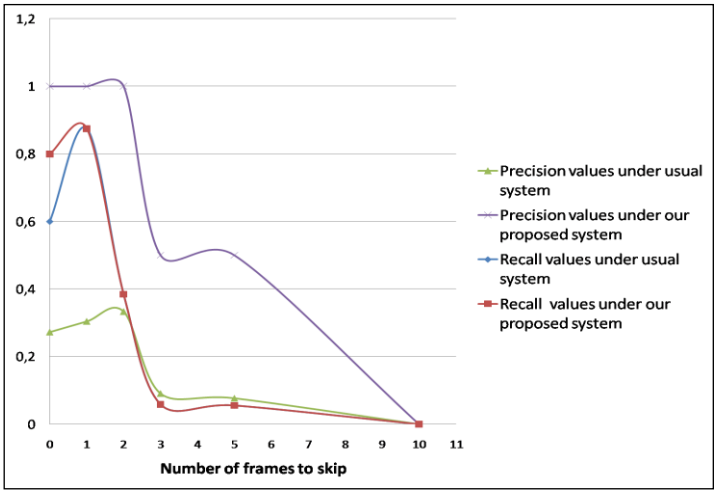

Figure 59. Comparaison des valeurs obtenues de rappel et de précision pour l'analyse de la séquence VISOR en variant le nombre d'images à ignorer avec un système de surveillance habituel et le système scalable orienté vers l'application

3.4. *Combinaison de la scalabilité orientée vers l'application avec la scalabilité spatio-temporelle*

La scalabilité orientée vers l'application est également évaluée en la combinant avec la scalabilité spatio-temporelle du codage vidéo. Les Figures 60, 61, 62 présentent respectivement les valeurs obtenues de PSNR, débit et temps de traitement des flux vidéos par les deux systèmes habituel et proposé en variant simultanément la résolution et la fréquence temporelle. Les mêmes conclusions sont extraites : les performances applicative, perceptuelle et computationnelle de notre système sont meilleures et l'étape d'analyse est correctement remplie. Fusionner la scalabilité orientée vers l'application avec la scalabilité spatiale et temporelle n'a pas d'impact régressif sur les améliorations déjà obtenues. Par conséquent, la scalabilité développée cohabite bien avec le codage vidéo scalable.

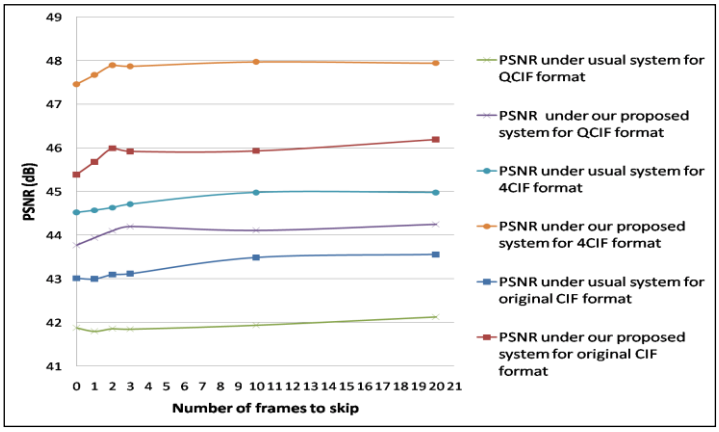

Figure 60. Comparaison des valeurs obtenues de PSNR pour la séquence VISOR en variant la résolution et la fréquence temporelle avec un système de surveillance habituel et le système scalable orienté vers l'application

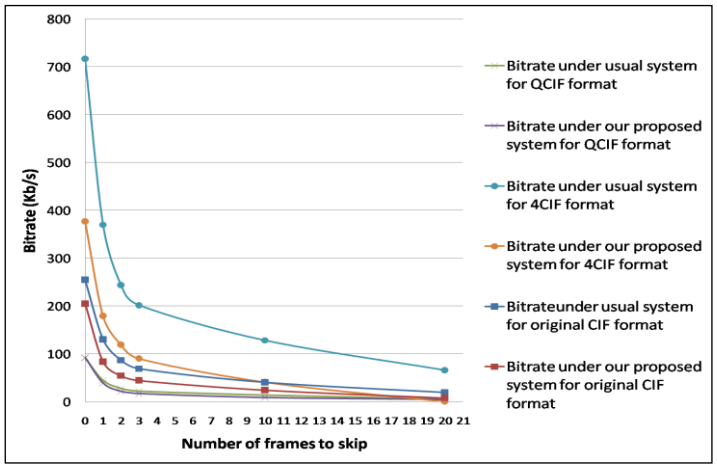

Figure 61. Comparaison des valeurs obtenues de débit binaire pour la séquence VISOR en variant la résolution et la fréquence temporelle avec un système de surveillance habituel et le système scalable orienté vers l'application

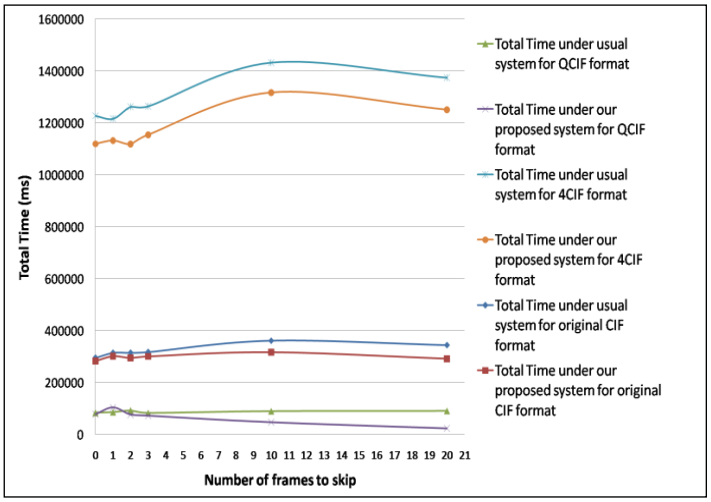

Figure 62. Comparaison des valeurs obtenues de temps d'exécution pour la séquence VISOR en variant la résolution et la fréquence temporelle avec un système de surveillance habituel et le système scalable orienté vers l'application

4. Conclusion

Nous avons présenté dans ce chapitre une architecture scalable orientée vers l'application pour les systèmes de vidéosurveillance. La scalabilité est appliquée sur l'architecture du système pour être capable d'atteindre les différents niveaux d'applications de surveillance. L'objectif général de ce travail est fournir une nouvelle architecture qui évite d'encourir les systèmes par le traitement de toutes les données enregistrées. Les informations pertinentes sont progressivement extraites, envoyées et traitées en fonction de l'application exigée par l'utilisateur final. Les résultats expérimentaux prouvent l'efficacité perceptuelle, computationnelle et applicative du système proposé par rapport au système ordinaire. La combinaison de la scalabilité orientée vers l'application avec la scalabilité spatiale et temporelle n'a aucun impact régressif sur les gains déjà obtenus. L'approche développée s'adapte avec le codage vidéo scalable. En même temps, l'étape d'analyse est réalisée avec succès. Il convient de dire, à la fin, que le système scalable développé n'est qu'un exemple proposé pour valider l'approche, et ainsi, beaucoup d'autres systèmes scalables orientés vers l'application peuvent être conçus.

Conclusion et perspectives

La vidéosurveillance reçoit, de plus en plus, beaucoup d'attention en tant qu'un domaine de recherche actif. La plupart des systèmes actuels peuvent traiter les flux vidéo et mettre en œuvre des fonctions analytiques de bas niveau. Récemment, l'intérêt technique de la vidéosurveillance a passé des tâches de bas niveau vers l'analyse des scènes plus complexes pour comprendre les comportements et interpréter les activités. L'analyse finale reste l'étape la plus importante dans tous les systèmes de vidéosurveillance. Elle est très dépendante de l'application souhaitée par l'utilisateur final. Cette phase est très discriminante, puisque toutes les applications de vidéosurveillance sont réalisées à travers une sélection de ses tâches exécutées d'une manière ascendante. Par conséquent, nous avons présenté une classification des applications des systèmes de vidéosurveillance en quatre catégories allant des applications de bas niveau vers des applications de haut niveau suivant les fonctions analytiques utilisées. La catégorie des applications de premier niveau regroupe les systèmes de surveillance qui détectent et comptent les objets dans la scène. Les systèmes appartenant au deuxième niveau applicatif sont capables de suivre les objets en mouvement. La catégorie des applications de troisième niveau peut classer et identifier les objets surveillés. La dernière classe comprend les systèmes qui analysent les comportements des objets en mouvement dans la scène. Décider l'application du système permet d'identifier les tâches à exécuter au cours de la phase d'analyse, qui, à son tour, définit les informations nécessaires du flux vidéo.

A travers les systèmes de vidéosurveillance existants, nous avons constaté que ces derniers encodent, transmettent et analysent toutes les données enregistrées tandis que réellement une faible partie des scènes est utile pour l'analyse. Ainsi, nous avons étendu l'architecture ordinaire des systèmes de surveillance par une phase de pré-analyse qui extrait les informations pertinentes pour l'analyse. L'objectif de cette étape est de détecter les régions d'intérêt dans la séquence et de les simplifier sans détruire les caractéristiques d'intérêt qu'elles contiennent pour les transmettre, ensuite, vers une analyse finale plus précise. Dans ce contexte, nous avons proposé deux méthodes différentes pour la pré-analyse dans le contexte de la vidéosurveillance. Une première consiste en un filtrage spatio-temporel des objets en mouvement de la scène. Elle simplifie les zones de l'image et ne garde que les informations intéressantes pour l'analyse afin de réduire les coûts matériels et logiciels du système de surveillance. Les évaluations expérimentales ont prouvé que la qualité visuelle des séquences

filtrées est suffisante et l'analyse des objets simplifiés de ce fait est bien réussie. En même temps, cette méthode maximise les redondances spatio-temporelles de la scène. Par conséquent, l'étape d'encodage H.264/AVC est améliorée en provoquant moins de dégradation à la qualité perceptuelle des images de la séquence. Également, les valeurs des erreurs de prédiction, le débit binaire et le temps d'exécution sont réduits. La deuxième méthode est une technique géométrique de modélisation de la forme qui simplifie la représentation des objets en mouvement en une forme parallélépipédique dans le plan de l'image. Les caractéristiques requises pour le suivi et la classification des objets sont maintenues. Cette méthode de modélisation est indépendante des paramètres intrinsèques de la caméra et de l'orientation de l'objet.

Jusqu'à présent, les systèmes de vidéosurveillance ont un seul niveau applicatif : chaque système est conçu pour répondre à une application spécifique. Nous avons proposé l'introduction du concept de la scalabilité orientée vers l'application au travers d'une architecture multi-niveaux applicatifs pour les systèmes de surveillance. Cette approche gère la scalabilité de point de vue applicatif. Elle affecte le niveau d'application en permettant à tout système de surveillance vidéo d'être adaptable et progressive relativement à sa réponse à l'utilisateur final. Le système peut être mis à jour, à la demande de l'utilisateur, pour atteindre des niveaux plus élevés. Les quatre niveaux d'applications cités peuvent être atteints grâce à un seul système. Cette architecture a une structure hiérarchique et fonctionne d'une manière scalable pour s'accorder avec les besoins progressifs de l'utilisateur final. Un exemple de système de vidéosurveillance respectant cette architecture et se basant sur nos méthodes de pré-analyse est proposé. C'est un système à quatre niveaux applicatifs où pour chaque niveau, les caractéristiques demandées sont extraites, envoyées et analysées en fonction de l'application souhaitée. En le comparant à un système de surveillance habituel, les résultats expérimentaux démontrent l'efficacité applicative, perceptuelle et computationnelle du système scalable. La scalabilité orientée vers l'application a été aussi combinée avec la scalabilité spatiale et temporelle du codage vidéo. Aucun impact régressif sur les gains déjà obtenus n'a été constaté et donc les deux types de scalabilité s'adaptent bien. Ainsi, nous pouvons exploiter les avantages de la scalabilité du codage vidéo et de la scalabilité orientée vers l'application dans un seul système de vidéosurveillance.

En conclusion, la créativité de cette thèse réside dans trois points : (1) La classification des systèmes de vidéosurveillance suivant le niveau de l'application; (2) L'extension de l'architecture habituelle des systèmes de vidéosurveillance par une étape de pré-analyse pour

ne garder que les informations pertinentes de la scène. (3) L'application du concept de la scalabilité dans l'aspect applicatif des systèmes de surveillance.

Parmi les perspectives de ces travaux, nous proposons d'améliorer le filtrage spatio-temporel en une méthode intelligente adaptative qui ajuste sa façon de simplifier les objets suivant des critères reliées à la scène et à la région détectée tels que: l'indice de texture, la distance par rapport à la caméra dans l'image, la nature et la taille de l'objet en mouvement. Il serait intéressant, aussi, de développer une méthode de pré-analyse appropriée pour l'interprétation des comportements qui réduit le taux d'information vidéo à transmettre et les coûts computationnels du système de surveillance tout en conservant les caractéristiques nécessaires pour réussir l'analyse. Comme autres perspectives, nous envisageons exploiter d'avantage l'architecture scalable orientée vers l'application et sa combinaison avec toutes les dimensions du codage vidéo scalable et proposer d'autres systèmes de vidéosurveillance où coexistent les deux types de scalabilité.

Publications de l'auteur

Ben Hamida, A.; Koubaa, M.; Nicolas, H.; Amar, C. "Video surveillance system based on a scalable application-oriented architecture", published online on Multimedia Tools and Applications, Springer US, 2015, 1-27.

Ben Hamida, A.; Koubaa, M.; Nicolas, H. ; Ben Amar, C. "Parallelepipedic shape modeling for moving objects in video surveillance systems", IEEE Science and Information Conference (SAI) 2014, vol., no., pp.379-383, 27-29 August, London, UK.

Ben Hamida, A.; Koubaa, M.; Nicolas, H. ; Ben Amar, C. "Toward Scalable application-oriented video surveillance systems", IEEE Science and Information Conference (SAI) 2014, vol., no., pp.384-388, 27-29 August, London, UK.

Ben Hamida, A.; Koubaa, M.; Nicolas, H. ; Ben Amar, C. "Video pre-analyzing and coding in the context of video surveillance", Multimedia and Expo (ICME), 2013 IEEE International Conference on , vol., no., pp.1,4, San Jose, USA, 15-19 July 2013.

Ben Hamida, A.; Koubaa, M.; Nicolas, H. ; Ben Amar, C. "Spatio-temporal video filtering for video surveillance applications", Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on , vol., no., pp.1,6, San Jose, USA, 19 July 2013.

Bibliographie

[Alayed 2013] H. Alayed, F. Frangoudes, C. Neuman, Behavioral-based cheating detection in online first person shooters using machine learning techniques, *In: IEEE Conference on Computational Intelligence in Games (CIG)*, pp 1–8, 2013.

[Albusac 2011] J. Albusac, J. Castro-Schez, D. Vallejo, L. Jimonez-Linares, and C. Glez-Morcillo, A scalable approach based on normality components for intelligent surveillance, In Innovations in Defence Support Systems, volume 336 of Studies in Computational Intelligence, pages 105-145, 2011.

[Ali 2001] A. Ali and J. Aggrawal, Segmentation and recognition of continuous human activity, *In IEEE Workshop on Detection and Recognition of Events in Video*, p. 28–35, 2001.

[Amato 2011] A. Amato, VD. Lecce, Semantic classification of human behaviors in video surveillance systems, *WSEAS Trans Comput* 10(10):343–352, 2011.

[Andriluka 2008] M. Andriluka, S. Roth and B. Schiele, People-tracking-by-detection and people-detection-by-tracking, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[Antonio 2013] C. Antonio, FL. David, SM. Antonio, JP. Juan, LD. María, Abandoned object detection on controlled scenes using kinect, *Nat Artif Comput Eng Med Appl* 7931:169–178, 2013.

[Aslani 2013] S. Aslani, H. Mahdavi-Nasab, Optical Flow Based Moving Object Detection and Tracking for Traffic Surveillance, *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, World Academy of Science, Engineering and Technology, 2013, 7, 786-790.

[Avidan 2004] S. Avidan, Support vector tracking, Pattern *Analysis and Machine Intelligence, IEEE transactions*, vol.26, (2004), pp.1064-1072.

[Babenko 2011] B. Babenko, M. H. Yang and S. Belongie, Robust object tracking with online multiple instance learning, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.33, (2011), pp.1619-1632.

[Bai 2011] YW. Bai, ZL. Xie, ZH. Li, Design and implementation of a home embedded surveillance system with ultra-low alert power, *IEEE Trans Consum Electron*, vol.57, 153–159, 2011.

[Barjatya 2004] A. barjatya, Block matching algorithms for motion estimation, Final project paper, DIP, Utah state university, 2004.

[Bektuzun 2013] E. Bektuzun, YS. Kucuksoz, KM. Elif, Real time tracking and detection of unusual circumstances of elderly people with RGB-d camera, *In: Signal Processing and Communications Applications Conference (SIU)*, 2013, pp 1–5.

[Bobick 2001] AF. Bobick, JW. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, 2001, 23, 257-267.

[Bondi 2000] A.B. Bondi, Characteristics of scalability and their impact on performance, *In Proceedings of the 2Nd International Workshop on Software and Performance, WOSP '00*, pages 195-203, New York, NY, USA, 2000.

[Bosch 2001] Bosch, Athens international airport, 2001, http://www.boschsecurity.co.uk/.

[Boulay 2006] B. Boulay, F. Bremond, M. Thonnat, Applying 3D human model in a posture recognition system, pattern recognition letter, Special Issue vis Crime Detect Prev. 27(15), 1788–1796, 2006.

[Boult 1999] T. Boult , Frame-rate omnidirectional surveillance and tracking of camuaged and occluded targets, *in Proceedings Second IEEE Workshop on Visual Surveillance*, pp. 48-55, Colorado, June 1999.

[Bradski 1998] G. Bradski, Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal, vol.2, no.2, pp.1-15, 1998.

[Brigger 2000] P. Brigger, J. Hoeg and M. Unser, B-spline snakes: a flexible tool for parametric contour detection, *Image Processing, IEEE Transactions*, vol.9, (2000), pp.1484-1496.

[Brulin 2012] M. Brulin, H. Nicolas, C. Maillet, Analyse d'un trafic routier dans un contexte de vidéo surveillance, *In: Proceedings of Sciences of Electronic, Technologies of Information and Telecommunications (SETIT)*, 2012.

[Brutzer 2011] S. Brutzer, B. Hoferlin, G. Heidemann, Evaluation of background subtraction techniques for video surveillance, *in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, vol., no., pp.1937-1944, 20-25 June 2011.

[BSIA 2013] British Security Industry Association, CCTV, 2013, http://www.bsia.co.uk/sections/cctv.aspx.

[Byung 2004] C.S Byung and W.C Kang. Motion-compensated temporal prefiltering for noise reduction in a video encoder, *In Image Processing, ICIP '04, 2004 International Conference on*, volume 2, pages 1221-1224 Vol.2, Oct 2004.

[Calderara 2006] S. Calderara, R. Melli, A. Prati, and R. Cucchiara, Reliable background suppression for complex scenes, in *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, ser. VSSN '06*, New York, NY, USA: ACM, 2006, pp. 211–214.

[Campbell 1995] L. Campbell and A. Bobick, Recognition of Human Body Motion Using Phase Space Constraints, *Proc. Int'l Conf. Computer Vision*, pp. 624-630, 1995.

[Cannons 2014] K. J. Cannons and R. P. Wildes, The applicability of spatiotemporal oriented energy features to region tracking, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.36, (2014), pp.784-796.

[Cavallaro 2004] A. Cavallaro, O. Steiger, and T. Ebrahimi, Perceptual prefiltering for video coding, *In Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pages 510-513, 2004.

[Chamasemani 2013] FF. Chamasemani, LS. Affendey, Systematic review and classification on video surveillance systems, *Int J Inf Technol Comput Sci* 5(7), pp: 87–102, 2013.

[Chang 2005] C. Chang, R. Ansari, and A. Khokhar, Multiple object tracking with kernel particle filter, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 566–573, 2005.

[Charpiat 2007] G. Charpiat, O. Faugeras and R. Keriven, Shape statistics for image segmentation with prior, *Computer Vision and Pattern Recognition, CVPR, IEEE Conference*, (2007).

[Chen 2005] T. Chen, H. Haussecker, A. Bovyrin, R. Belenov, K. Rodyushkin, A. Kuranov, V. Eruhimov, Computer vision workload analysis: case study of video surveillance systems, Intell Technol J 9(2).

[Chen 2013] YL. Chen, TS. Chen, TW. Huang, LC. Yin, SY. Wang, T. Chiueh, Intelligent urban video Surveillance system for automatic vehicle detection and tracking in clouds, *Advanced Information Networking and Applications , 2013 IEEE 27th International Conference on* , pp.814,821, 25-28 March.

[Cheung 2004] SC. Cheung, C. Kamath, Robust techniques for background subtraction in urban traffic video, *Visual Communications and Image Processing 2004*, 5308, 881-892.

[Cheung 2005] SC Cheung and C. Kamath, Robust background subtraction with foreground validation for urban traffic video, *EURASIP J. Appl. Signal Process*., vol. 2005, pp. 2330–2340, Jan. 2005.

[Ching 2003] C. Shu-Ching, S. Mei-Ling, S. Peeta, and Z. Chengcui, Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database systems, *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 3, pp.154–167, 2003.

[Chua 2015] JL. Chua, YC. Chang, WK. Lim, A simple vision-based fall detection technique for indoor video surveillance, *Signal Image and Video Processing*, 9(3):623–633, 2015.

[Chundi 2015] M. Chundi, X. Jianbin, Y. Wei, L.Tong, L. Peiqin, A fast recognition algorithm for suspicious behavior in high definition videos, *Multimed Syst* , 2015, 1-11.

[Collins 2000] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, A system for video surveillance and monitoring, VSAM final report[R], Carnegie Mellon University Technical Report CMU-RI-TR-00-12, 2000.

[Comaniciu 2000] D. Comaniciu, V. Ramesh and P. Meer, Real-time tracking of non-rigid objects using mean shift, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.142-149, 2000.

[Comaniciu 2002] D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.24, (2002), pp.603-619.

[Comaniciu 2003] D. Comaniciu, V. Ramesh and P. Meer, Kernel-based object tracking, *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol.25, no.5, pp.564-575, 2003.

[Coppi 2011] D. Coppi, S. Calderara, R. Cucchiara, Iterative active querying for surveillance data retrieval in crime detection and forensics, *In: 4th International Conference on Imaging for Crime Detection and Prevention (ICDP)*, 2011, pp 1–6.

[Corporation 2009] Corporation I, Command, control, collabo-rate: public safety solutions from IBM, 2009, Solution Brief.

[Cucchiara 2001] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, Improving Shadow Suppression in Moving Object Detection with HSV Color Information, *Proc. IEEE Int'l Conf. Intelligent Transportation Systems*, pp. 334-339, Aug. 2001.

[Grana 2001] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, Detecting objects, shadows and ghosts in video streams by exploiting color and motion information, *in Image Analysis and Processing, 2001. Proceedings. 11th International Conference on*, Sep 2001, pp. 360–365.

[Cucchiara 2005] R. Cucchiara, A. Prati, R. Vezzani, Posture classification in a multicamera indoor environment, *in Proceedings of IEEE International Conference on Image Processing (ICIP)*, 1, pp. 725-728, Genova, Italy, September 2005.

[Cui 1995] Y. Cui, D. Swets, and J. Weng, Learning-Based Hand Sign Recognition Using Shoslif-m, *Proc. Int'l Conf. Computer Vision*, pp. 631-636, 1995.

[Cupillard 2001] F. Cupillard, F. Brémond, M. Thonnat, Tracking groups of people for video surveillance, *in Proceedings of the European Workshop on Advanced Video Based Surveillance Systems (AVBSS01)*, Kingston, United Kingdom, 2001.

[Dahlkamp 2007] H. Dahlkamp, HH. Nagel, A. Ottlik and P. Reuter, A framework for model-based tracking experiments in image sequences, *International Journal of Computer Vision*, vol.73, (2007), pp.139-157.

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

[Darrell 1993] T. Darrell and A. Pentland, Space-Time Gestures, *Proc. Computer Vision and Pattern Recognition*, pp. 335-340, 1993.

[Dee 2008] H.M. Dee and S.A. Velastin, How close are we to solving the problem of automated visual surveillance, *Machine Vision and Applications*, 19((5- 6)):329-343, 2008.

[Derrode 2006] S. Derrode, M. A. Charmi and F. Ghorbel, Fourier-based invariant shape prior for snakes, *Acoustics, Speech and Signal Processing, ICASSP Proceedings. IEEE International Conference*, (2006).

[Detmold 2006] H. Detmold, A. Dick, K. Falkner, D.S. Munro, A. van den Hengel, and R. Morrison, Scalable surveillance software architecture, *In Video and Signal Based Surveillance, 2006. AVSS '06. IEEE International Conference on*, pages 103-103, Nov 2006.

[Dimitropoulos 2009] K. Dimitropoulos, T. Semertzidis, N. Grammalidis, Video and signal based surveillance for airport applications, *In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) IEEE*, 2009, pp 170–175.

[Dongbin 2014] LV. Le, Z. Dongbin, F. Zhijiang, Cheating behavior detection based-on pictorial structure model, *In: 33rd Chinese Control Conference (CCC)*, 2014, pp 7274–7279.

[Duman 2013] D. Duman, GB. Akar, Moving vehicle classification, *In : Conference on Signal Processing and Communications Applications (SIU)*, 2013, pp 1–4.

[Elgammal 1999] A. Elgammal, D. Harwood, and L. Davis, Non-parametric model for background subtraction, *in Proceedings of IEEE ICCV'99 Frame-rate workshop*, Sept 1999.

[Essa 1997] I. Essa and A. Pentland, Coding, Analysis, Interpretation, and Recognition of Facial Expressions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757-763, July 1997.

[Ezzahout 2013] A. Ezzahout, ROH. Thami, Conception and development of a video surveillance system for detecting, tracking and profile analysis of a person. *In: 3rd International Symposium for Knowledge Organization (ISKO)*, 2013, pp 1–5.

[Fang 2011] H. Fang, S. H. Kim and J. W. Jang, A snake algorithm for automatically tracking multiple objects, *Image Processing (ICIP), 18th IEEE International Conference*, 2011.

[Farlane 1995] MC. Farlane and C. Schoffield, Segmentation and tracking of piglets in images, *Machine Vision and Applications* 8(3), pp. 187-193, 1995.

[Feng 2012] L. Chia-Feng, Y. Shyan-Ming, L. Muh-Chyi, and T. Ching-Tsorng, A framework for scalable cloud video recorder system in surveillance environment, *In Ubiquitous Intelligence Computing and 9th International Conference on Autonomic Trusted Computing (UIC/ATC), 2012 9th International Conference on*, pages 655-660, Sept 2012.

[Fuhai 2012] W. Lei, L. Fuhai, The design of real-time monitoring system for enterprises in complex environments, *In: International Conference on Systems and Informatics (ICSAI)*, 2012, pp 306–314.

[Gavrila 1996] D. M. Gavrila and L. S. Davis, 3-D model-based tracking of humans in action: A multi-view approach, *Computer Vision and Pattern Recognition, Proceedings CVPR, IEEE Computer Society Conference*, (1996).

[Gong 2011] S. Gong, C. Loy, and T. Xiang, Security and surveillance, *In Visual Analysis of Humans*, 455-472, Springer London, 2011.

[Gordon 1993] N. J. Gordon, D. J. Salmond and A. F. M. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings F (Radar and Signal Processing), IET Digital Library*, (1993).

[Gouaillier 2009] V. Gouaillier and A.E. Fleurant, Intelligent video surveillance: Promises and challenges technological and commercial intelligence report, Technical report, CRIM and Technôpole Defence and Security, 2009.

[Grois 2010] D. Grois, E. Kaminsky, and O. Hadar, Roi adaptive scalable video coding for limited bandwidth wireless networks, In Wireless Days (WD), 2010 IFIP, pages 1-5, 2010.

[Grzegorz 2014] S. Grzegorz, D. Piotr, Detection of vehicles stopping in restricted zones in video from surveillance cameras, Multimed Commun Serv Secur 429:242–253, 2014.

[Hager 1998] G. D. Hager and P. N. Belhumeur, Efficient region tracking with parametric models of geometry and illumination, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.20, (1998), pp.1025-1039.

[Halevy 1999] G. Halevy and D. Weinshall, Motion of disturbances: detection and tracking of multi-body non-rigid motion, *Maching Vision and Applications* 11, pp. 122-137, 1999.

[Haritaoglu 2000] I. Haritaoglu, D. Harwood, and L. S. Davis, W4: Real-time surveillance of people and their activities, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no.8, pp.809–830, 2000.

[He 2014] R. He, B. Yang and N. Sang, Integral region-based covariance tracking with occlusion detection, *Multimedia Tools and Applications*, (2014), pp.1-22.

[Heckbert 1990] P. Heckbert, A Seed Fill Algorithm (Graphics Gems I), New York: Academic Press, 1990.

[Héctor 2015] F. Héctor, A. Gómez, MT. Rafael, AT. Susana, FC. Antonio, R. Sylvie, GE. Alexandra, LG. Patricia, Identification of loitering human behaviour in video surveillance environments, *In: International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC)*, 2015, pp 516–525.

[Heikkila 1999] J. Heikkila and O. Silven, A real-time system for monitoring of cyclists and pedestrians, *in Second IEEE Workshop on Visual Surveillance*, pp. 246-252, Colorado, June 1999.

[Hill 1990] M.D. Hill, What is scalability?, SIGARCH Comput. Archit. News, 18(4):18-21, December 1990.

[Hitesh 2013] AP. Hitesh, GT. Darshak, Moving Object Tracking Using Kalman Filter, *International Journal of Computer Science and Mobile Computing*, April 2013, pg.326 – 332.

[Horn 1981] BKP. Horn, BG. Schunck, Determining optical flow, *Artificial Intelligence*, vol.17, pp 185–203, 1981.

[Hosik 2009] S. Hosik, E.T. AnzaKu, W. De Neve, R. Yong-Man, and K.N. Plataniotis, Privacy protection in video surveillance systems using scalable video coding, *In Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pages 424-429, 2009.

[Hota 2007] R. N. Hota, V. Venkoparao and A. Rajagopal, Shape based object classification for automated video surveillance with feature selection, *IEEE, 10th International Conference on Information Technology*, Orissa, pp. 97-99 Dec 2007.

[Houhou 2008] N. Houhou, J. Thiran and X. Bresson, Fast texture segmentation model based on the shape operator and active contour, *Computer Vision and Pattern Recognition, CVPR, IEEE Conference*, (2008).

[Hu 2004] W. Hu, T. Tan, L. Wang et S. Maybank, A survey on visual surveillance of object motion and behaviors, *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, vol. 34, no. 3, pages 334-352, 2004.

[Ibrahim 2012] AM. Ibrahim, AA. Shafie, MM. Rashid, Human identification system based on moment invariant features, *Computer and Communication Engineering, 2012 International Conference on*,pp.216-221, 3-5 July 2012.

[IHS 2014] IHS Technology (NYSE: IHS),Trends for 2014 - Video Surveillance Trends for the Year Ahead.

[Imran 2013] S. Imran, S. Mubarak, Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos, *Int J Comput Vis* 104(2):198–219, 2013.

[Jalal 2012] AS. Jalal, The State-of-the-Art in Visual Object Tracking, Informatica, 2012, 36, 227-248.

[Javed 2002] O. Javed and M. Shah, Tracking and object classification for automated surveillance, *proceedings of the 7th European Conference on Computer Vision-Part IV*, Springer-Verlag, 2002, 343-357.

[Javed 2003] O. Javed, R. Zeeshan, O. Alatas, et al, Knight M : A real time surveillance system for multiple overlapping and non-overlapping cameras [J], *The fourth International Conference on Multimedia and Expo*, Baltimore, Maryland, 2003.

[Jepson 2003] A. D. Jepson, D. J. Fleet and T. F. El-Maraghi, Robust online appearance models for visual tracking, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.25, (2003), pp.1296-1311.

[Johnsen 2009] S. Johnsen and A. Tews, Real-time object tracking and classification using a static camera, *Proc. of the IEEE ICRA 2009 Workshop on People Detection and Tracking*, Japan, May 2009.

[Jun 2015] L. Jun L, W. Jinqiao, X. Huazhong, L. Hanqing, A real-time people counting approach in indoor environment, *Multimed Model* 8935:214–223, 2015.

[Kalal 2012] Z. Kalal, K. Mikolajczyk and J. Matas, Tracking-Learning-Detection, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.34, (2012), pp.1409-1422.

[Kalman 1960] R. E. Kalman, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, vol.82, (1960), pp.35-45.

[Kameda 1996] Y. Kameda and M. Minoh, A human motion estimation method using 3-successive video frames, *In ICVSM*, pages 135–140, 1996.

[Karaca 2000] H. M. Karaca, E. Anarm, and A. Morgul, Role of prefiltering in unsupervised video segmentation, *In Proceedings of the Acoustics, Speech, and Signal Processing, 2000. On IEEE International Conference - Volume 04, ICASSP '00*, pages 1999-2002, Washington, DC, USA, 2000.

[Karaulova 2002] I. A. Karaulova, P. M. Hall and A. D. Marshall, Tracking people in three dimensions using a hierarchical model of dynamics, *Image and Vision Computing*, vol.20, (2002), pp.691-700.

[Karmann 1990] K.-P. Karmann and A. Brandt, Moving object recognition using and adaptive background memory, in Time-Varying Image Processing and Moving Object Recognition, V. Cappellini, ed., 2, pp. 289{307, Elsevier Science Publishers B.V.,1990.

[Kass 1988] M. Kass, A. Witkin and D. Terzopoulos, Snakes: Active contour models , *International Journal of Computer Vision*, vol.1, (1988), pp.321-331.

[Khan 2009] S. M. Khan and M. Shah, Tracking Multiple Occluding People by Localizing on Multiple Scene Planes, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.31, (2009), pp.505-519.

[Ko 2008] T. Ko, A survey on behavior analysis in video surveillance for homeland security applications, *In: 37th Workshop on Applied Imagery Pattern Recognition (AIPR) IEEE*, 2008, pp 1–8.

[Koller 1993] D. Koller, J. Weber, and J. Malik, Robust multiple car tracking with occlusion reasoning, Tech. Rep. UCB/CSD-93-780, EECS Department, University of California, Berkeley, Oct 1993.

[Kruger 2012] M. Kruger, L. Ziegler, K. Heller, A generic Bayesian Network for identification and assessment of objects in maritime surveillance, *In:15th International Conference on Information Fusion (FUSION)*, 2012, pp 2309–2316.

[Lai 2007] CL. Lai, JC. Yang, YH. Chen, A real time video processing based surveillance system for early fire and flood detection, *In: Instrumentation and Measurement Technology Conference Proceedings (IMTC)*, 2007, pp 1-6.

[Lambert 2006] P. Lambert, D. De Schrijver, D. Van Deursen, W. De Neve, Y. Dhondt, and R. Van de Walle, A real-time content adaptation framework for exploiting roi scalability in h.264/avc, *In Proceedings of the 8th international conference on Advanced Concepts For Intelligent Vision Systems, ACIVS'06*, pages 442-453, Berlin, Heidelberg, 2006.

[Le 2012] A. Le, M. Kafai, B. Bhanu, Face recognition in multi-camera surveillance videos using Dynamic Bayesian Network, *In: Sixth International Conference on Distributed Smart Cameras (ICDSC)*, pp1-6,2012.

[Le 2014] A. Le, B. Bir, Y. Songfan, Unified face representation for individual recognition in surveillance videos, Wide Area Surveill 6:123–136, 2014.

[Lee 2009] P-H. Lee, T-H. Chiu, Y-L. Lin, Y-P. Hung, Real-time pedestrian and vehicle detection in video using 3d cues, *in Proceedings of the 2009 IEEE international conference on Multimedia and Expo, ICME'09*, IEEE Press, Piscataway 614–617, 2009.

[Lee 2011] J.-Y Lee and W. Yu, Visual tracking by partition-based histogram back projection and maximum support criteria, *IEEE Conference on Robotics and Biomimetics*, pp. 2860-2865, Dec 2011.

[Leichter 2012] I. Leichter, Mean shift trackers with cross-bin metrics, *Pattern Analysis and Machine intelligence, IEEE Transactions*, vol.34, (2012), pp.695-706.

[Li 2003] L. Li, W. Huang, I.Y.H. Gu, and Q. Tian, 2003. Foreground object detection from videos containing complex background, In *Proceedings of the eleventh ACM international conference on Multimedia* (MULTIMEDIA '03). ACM, New York, NY, USA, 2-10, 2003.

[Li 2004] A. Yilmaz, X. Li and M. Shah, Object contour tracking using level sets, *Asian Conference on Computer Vision*, (2004).

[Li 2012] G. R. Li, Q. M. Huang, J. B. Pang, S. Q. Jiang and L. Qin, Online selection of the best feature subset for object tracking, *Journal of Visual Communication and Image Representation*, vol.23, (2012), pp.254-263.

[Little 1995] J. Little and J. Boyd, Describing Motion for Recognition, *Int'l Symp. Computer Vision*, pp. 235-240, Nov. 1995.

[Lo 2001] B. P. L. Lo and S. Velastin, Automatic congestion detection system for underground platforms, *in Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, 2001, pp. 158–161.

[Loomans 2011] MJH. Loomans, JC. Koeleman, Low-complexity wavelet-based scalable image & video coding for home-use surveillance, *IEEE Trans Consum Electron*, vol.57, 507–15, 2011.

[Lucas 1981] B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, *IJCAI*, vol.81, (1981), pp.674-679.

[Luo 2011] X. Luo, Y. Wu, Y. Huang, J. Zhang, Vehicle flow detection in real-time airborne traffic surveillance system, *Trans Inst Meas Control* 33:880–897, 2011.

[Maddalena 2013] L. Maddalena, A. Petrosino, Stopped object detection by learning foreground model in videos, *IEEE Transactions on Neural Networks and Learning Systems* (24)5:723–735, 2013.

[Mahalingam 2010] T. Mahalingam and M. Mahalakshmi, Vision based moving object tracking through enhanced color image segmentation using Haar classifiers, *IEEE Conference on Trendz in Information Sciences and Computing*, pp. 253-260, Dec 2010.

[Melli 2005] R. Cucchiara, R. Melli, A. Prati, L. De Cock, Predictive and probabilistic tracking to detect stopped vehicles, *in Proceedings of Workshop on Applications of Computer Vision (WACV)*, pp. 388.393, Breckenridge, USA, 4-7 January 2005.

[Melo 2006] J. Melo, A. Naftel, A. Bernardino, and J. Santos-Victor (2006), Detection and classification of highway lanes using vehicle motion trajectories, *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no.2, pp.188–200.

[Meng 2015] L. Meng, C. Zemin, W. Chuliang and Y. Ye , A Survey of Video Object Tracking, *International Journal of Control and Automation* Vol. 8, No. 9 (2015), pp. 303-312.

[Moeslund 2006] T.B. Moeslund, A. Hilton et V. Kruger, A survey of advances in vision based human motion capture and analysis, *Computer vision and image understanding*, vol. 104, no. 2-3, pages 90-126, 2006.

[Mukherjee 2011] S. Mukherjee, B. Saha, I. Jamal, R. Leclerc, N. Ray, A novel framework for automatic passenger counting, *In:18th IEEE International Conference on Image Processing*, 2011, pp 2969–2972.

[Najman 1994] L. Najman and M. Schmitt, Ligne de partage des eaux, In Michel Schmitt and Juliette Mattioli, editors, Morphologie Mathématique, pages 121-140. Masson, 1994.

[Negri 2014] P. Negri, Estimating the queue length at street intersections by using a movement feature space approach. *IET Image Process* 8(7):406–416, 2014.

[Norris 2004] C. Norris, M. McCahill, and D. Wood, Editorial. the growth of cctv: a global perspective on the international diffusion of video surveillance in publicly accessible space, Surveillance Society, Vol. 2, Issue 2/3, 2004.

[Nouar 2006] O. D. Nouar, G. Ali and C. Raphael, Improved object tracking with CamShift algorithm, *Acoustics, Speech and Signal Processing, ICASSP Proceedings, IEEE International Conference*, 2006.

[Novak 1981] L. M. Novak, Optimal target designation techniques, *IEEE Transactions on Aerospace and Electronic Systems*, vol. 17, no.5, pp. 676–684, 1981.

[Nummiaro 2003] k. Nummiaro, E.Koller-Meier, L.V Gool, An adaptive color-based particle filter, *Image and Vision Computing*, Volume 21, Issue 1, 10 January 2003, Pages 99-110.

[Palmares 2011] http://owni.fr/2011/12/15/le-palmares-des-villes-sous-surveillance/.

[Paragios 2002] N. Paragios and R. Deriche, Geodesic active regions and level set methods for supervised texture segmentation, *International Journal of Computer Vision*, vol.46, (2002), pp.223-247.

[Parekh 2014] H. S. Parekh, D. G. Thakore and U. K. Jaliya, A survey on object detection and tracking methods, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, Issue 2, Feb 2014.

[Peng 2014] Z. Peng, Z. Yanning, T. Tony, E. Sabu, Moving people tracking with detection by latent semantic analysis for visual surveillance applications, *Multimedia Tools and Applications*, 68(3):991–1021, 2014.

[Protection 2014] Protection Sécurité Magazine date : 31/01/2014.

[Ramanan 2003] D. Ramanan and D. A. Forsyth, Finding and tracking people from the bottom up, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 467–474.

[Rodgers 1988] J. L. Rodgers, and W. A. Nicewander, Thirteen ways to look at the correlation coefficient, American Statistician, vol. 42, p. 59-66, 1988.

[Ronfard 1994] R. Ronfard, Region-based strategies for active contour models, *International Journal of computer Vision*, vol.13, (1994), pp.229-251.

[Ross 2008] D. A. Ross, J. Lim, R. S. Lin and M. H. Yang, Incremental learning for robust visual tracking, *International Journal of Computer Vision*, vol.77, (2008), pp.125-141.

[Salvi 2014] G. Salvi, An automated nighttime vehicle counting and detection system for traffic surveillance, *In: International Conference on Computational Science and Computational Intelligence*, pp 131–136, 2014.

[Sanghyuk 2012] P. Sanghyuk, CD. Yoo, Video scene analysis and irregular behavior detection for intelligent surveillance system, *In: 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp 577–581, 2012.

[Schwarz 2007] H. Schwarz, D. Marpe, and T. Wiegand, Overview of the scalable video coding extension of the h.264/avc standard, *IEEE Trans. Cir. and Sys. for Video Technol*., 17(9):1103-1120, September 2007.

[Scotti 2005] G. Scotti, A. Cuocolo, C. Coelho, L. Marchesotti, A novel pedestrian classification algorithm for a high definition dual camera 360 degrees surveillance system, *in Proceedings of the International Conference on Image Processing (ICIP 2005)*, Genova, Italy. 3, 880–883, 2005.

[Sehchan 2007] O. Sehchan, P. Sunghyuk, L. Changmu, A platform surveillance monitoring system using image processing for passenger safety in railway station, *In: International Conference on Control, Automation and Systems, ICCAS*, 2007, pp 394–398.

[Senior 2006] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, R. Bolle, Appearance models for occlusion handling, *Image and Vision Computing*, Volume 24, Issue 11, 1 November 2006, Pages 1233-1243.

[Senior 2009] A. Senior, Protecting Privacy in Video Surveillance, Springer Publishing Company, Incorporated, 2009.

[Sergyn 2007] S. Sergyn, Color content-based image classification, *5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics*, Poprad, Slovakia, Jan 2007.

[Sethi 1987] I. K. Sethi and R. Jain, Finding trajectories of feature points in a monocular image sequence, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.1, (1987), pp.56-73.

[Setitra 2014] I. Setitra, S. Larabi, Background Subtraction Algorithms with Post-processing: A Review, *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014, pp: 2436-2441.

[Shan 2007] C. F. Shan, T. N. Tan and Y. C. Wei, Real-time hand tracking using a mean shift embedded particle filter, *Pattern Recognition*, vol.40, (2007), pp.1958-1970.

[Shizheng 2011] W. Shizheng, Y. Jianwei, Z. Yanyun, A. Cai, and S.Z. Li, A surveillance video analysis and storage scheme for scalable synopsis browsing, *In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1947-1954, Nov 2011.

[Simone 2014] C. Simone, M. Lucio, M. Pietro, R. Carlo, Event based switched dynamic bayesian networks for autonomous cognitive crowd monitoring, Wide Area Surveillance : pp 93–122, 2014.

[Sinha 2011] S. N. Sinha, J. M. Frahm, M. Pollefeys and Y. Genc, Feature tracking and matching in video using programmable graphics hardware, *Machine Vision and Applications*, vol.22, (2011), pp.207-217.

[Smith 2004] G.J.D Smith, Behind the screens : examining constructions of deviance and informal practices among cctv control room operators in the UK, Surveillance Society, 2(2/3):376-395, 2004.

[Stauffer 1999] C. Stauffer, and W. Grimson, Adaptive background mixture models for real-time tracking, *Computer Vision and Pattern Recognition*, vol. 2, pp. 246-252, 1999.

[Sung 2014] CL. Sung, N. Ram, Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system, *Mach Vis Appl* 25(1):133–143, 2014.

[Surveillance 2013]Video Surveillance and VSaaS Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast, 2013 – 2019.

[Suzuki 1985] S. Suzuki, K. Abe, Topological structural analysis of digitized binary images by border following, *CVGIP* 30 1, pp 32-46 (1985).

[Szpak 2011] ZL. Szpak, JR. Tapamo, Maritime surveillance: tracking ships inside a dynamic background using a fast level-set, *Expert Syst Appl* 38(6):6669–66680, 2011.

[Tissainayagam 2005] P. Tissainayagam and D. Suter, Object tracking in image sequences using point features, *Pattern Recognition*, vol.38, (2005), pp.105-113.

[Toyama 1999] K. Toyama, J. Krumm, B. Brumitt, B. Meyers. Wall_ower : Principles and practice of background maintenance. *In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 255-261, 1999.

[Tsai 2006] YT. Tsai, HC. Shih, and CL. Huang, Multiple human objects tracking in crowded scenes, *IEEE Conference on Pattern Recognition*, Vol. 3, pp. 51-54, 2006.

[Tsuji 2002] H. Tsuji, T. Sakatani, Y. Yashima, and N. Kobayashi, A nonlinear spatiotemporal diffusion and its application to prefiltering in mpeg-4 video coding. *In Image Processing, Proceedings. 2002 International Conference on*, volume 1, pages I-85-I-88 vol.1, 2002.

[Tsung 2012] C. Hua-Tsung, T. Li-Wu, G. Hui-Zhen, L. Suh-Yin, BSP. Lin, Traffic Congestion Classification for Nighttime Surveillance Videos, *In: IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012, pp 169–174.

[Tuty 2014] J. Tuty and B. Zhang, Simultaneous object tracking and classification for traffic surveillance, *In Proceedings of International Conference on Computer Science and Information Technology*, volume 255 of Advances in Intelligent Systems and Computing, pages 749-755, Springer India, 2014.

[Veenman 2001] C. Veenman, M. Reinders, E. Backer, Resolving motion correspondence for densely moving points, *IEEE Trans. Patt. Analy. Mach. Intell*., vol. 23, no. 1, pp. 54–72, 2001.

[Wallace 1988] E. Wallace and C. Diffley, Cctv control room ergonomics, Technical Report 14/98, Police Scientific Development Branch (PSDB), UK Home Office, 1988.

[Wan 2011] P. Su-Wan, W.H. Jong, and S. Sang-Uk, Secure service mechanism of video surveillance system based on h.264/svc, *In Information Technology and Multimedia (ICIM), 2011 International Conference on*, pages 1-4, Nov 2011.

[Wang 2012] Y. Wang, X. Su, M. Yang, L. Xu, C. Tang, A violations stop detect system based on surveillance camera, *Consumer Electronics, Communications and Networks, 2012 2nd International Conference on* ,pp.3086-3089, 21-23 April.

[Wang 2015] S. Wang, L. Chen, Z. Zhou, X. Sun, J. Dong, Human fall detection in surveillance video based on PCANet, *Multimedia Tools and Applications*, pp: 1–11, 2015.

[Wedi 2003] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra, Overview of the H.264/AVC video coding standard, *in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.

[Weiming 2004] H. Weiming, T. Tieniu, W. Liang, and S. Maybank, A survey on visual surveillance of object motion and behaviors, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3):334-352, Aug 2004.

[Williams 2005] O. Williams, A. Blake and R. Cipolla, Sparse bayesian learning for efficient visual tracking, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.27, (2005), pp.1292-1304.

[Wilson 1995] A. Wilson and A. Bobick, Learning Visual Behavior for Gesture Analysis, *Proc. IEEE Int'l. Symp. Computer Vision*, Nov 1995.

[Wren 1997] C. Wren, A. Azabayejani, T. Darrel, and A. Pentland, Pfinder: Real-time tracking of the human body, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, pp. 780-785, July 1997.

[Wu 2004] S. Wu, L. Hong and J. R. Layne, 2D rigid-body target modeling for tracking and identification with GMTI/HRR measurements, *Control Theory and Applications, IEE Proceedings*, vol.151, (2004), pp.429-438.

[Xinfeng 2014] B. Xinfeng, S. Javanbakhti, S. Zinger, R. Wijnhoven, Context-based object-of-interest detection for a generic traffic surveillance analysis system, *In: 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014, pp 136–141.

[Xue 2002] Z. Xue, S. Z. Li and E. K. Teoh, AI-EigenSnake: an affine-invariant deformable contour model for object matching, *Image and Vision Computing*, vol.20, (2002), pp.77-84.

[Yang 2001] H. Yang, J. Lou, H. Z. Sun, W. M. Hu and T. N. Tan, Efficient and robust vehicle localization, *Image Processing, Proceedings International Conference*, (2001).

[Yang 2005] C. Yang, R. Duraiswami, and L. Davis, Fast multiple object tracking via a hierarchical particle filter, *In Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 212–219, 2005.

[Yaning 2015] W. Yaning, Z. Hong, Pedestrian detection and counting based on ellipse fitting and object motion continuity for video data analysis, *Intell Comput Theor Methodologies* 9225:378–387, 2015.

[Yilmaz 2004] A. Yilmaz, X. Li and M. Shah, Contour based object tracking with occlusion handling in video acquired using mobile cameras, *IEEE Trans. Patt. Analy. Mach. Intell*. vol.26, no. 11, pp. 1531–1536, 2004.

[Yilmaz 2006] A. Yilmaz, O. Javed and M. Shah (2006). Object tracking: a survey, *ACM Journal of Computing Surveys*, vol. 38, no.4, Article 13.

[Yoneyama 2005] A. Yoneyama, C. Yeh, C. Kuo, Robust vehicle and traffic information extraction for highway surveillance, *EURASIP J Appl Signal Process*, 2305–2321, 2005.

[Yongil 2010] C. Yongil, O.L. Sang, and H.S. Yang, Collaborative occupancy reasoning in visual sensor network for scalable smart video surveillance, *Consumer Electronics, IEEE Transactions on*, 56(3):1997-2003, Aug 2010.

[Yun 2014] K. Yun, H. Jeong, KM. Yi, SW. Kim, JY. Choi, Motion interaction field for accident detection in traffic surveillance video, *In: 22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp 3062–3067.

[Zang 2003] Q. Zang, R. Klette, Object classification and tracking in video surveillance, *Computer Analysis of Images and Patterns: 10th International Conference, CAIP 2003*, p. 198-205, Groningen, The Netherlands, August 25-27, 2003.

[Zhong 2000] Y. Zhong, A. K. Jain and M. P. Dubuisson-Jolly, Object tracking using deformable templates, *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol.22, (2000), pp.544-549.

[Zhou 2004] W. Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *Image Processing, IEEE Transactions on*, 13(4):600-612, April 2004.

[Zhou 2006] Q. M. Zhou and J. K. Aggarwal, Object tracking in an outdoor environment using fusion of features and cameras, *Image and Vision Computing*, vol.24, (2006), pp.1244-1255.

[Zhuo 2013] W. Zhuo, R.H. Deng, S. Jialie, W. Yongdong, D. Xuhua, and L. SweeWon, Technique for authenticating h.264/svc streams in surveillance applications, *In Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1-4, July 2013.

[Ziliani 2003] F. Ziliani and J. Reichel, Second generation prefiltering for video compression and analysis in multisensor surveillance systems, *In Multisensor Surveillance Systems*, pages 119-133, Springer US, 2003.

[Ziliani 2005] F. Ziliani, The importance of 'scalability' in video surveillance architectures, *In Imaging for Crime Detection and Prevention, 2005. ICDP 2005. The IEE International Symposium on*, pages 29-32, 2005.

[Zivkovic 2004] Z. Zivkovic , Improved adaptive Gaussian mixture model for background subtraction, *International Conference Pattern Recognition*, vol. 2, p. 28-31, 2004.

[Zivkovic 2006] Z. Zivkovic, and F. van der Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, *Pattern Recognition Letters*, vol. 27, p. 773-780, 2006.

[Zoidi 2013] O. Zoidi, A. Tefas and I. Pitas, Visual object tracking based on local steering kernels and color histograms, *Circuits and Systems for Video Technology, IEEE Transactions*, vol.23, (2013), pp.870-882.

[Zuniga 2006] M. Zuniga, F. Bremond, and M. Thonnat, Fast and reliable object classification in video based on a 3d generic model, *in Proceedings of the International Conference on Visual Information Engineering (VIE2006)*, (Bangalore, India), pp. 433–440, 26-28, September 2006.

[Zuniga 2011] M. Zuniga, F. Bremond, and M. Thonnat, Uncertainty control for reliable video understanding on complex environments, in Video Surveillance (W. Lin, ed.), ch. 21, pp. 383-408, INTECH, February 2011.