Liste des acronymes

OVIS: *Ontology Video-Surveillance Indexing and Retrieval System.*
SWRL: *Semantic Web Rule Language.*
MJPEG: *Motion Joint Photographic Experts Group.*
MPEG: *Moving Picture Experts Group.*
GSM: *Global System for Mobile Communications.*
PDA: *Personal Digital Assistant.*
CD-ROM: *Compact Disc - Read Only Memory.*
DVD : *Digital Versatile Disc.*
DOC : *Document.*
ODT : *Open Document Text.*
ZIP : *Compressed File.*
PDF : *Portable Document Format.*
RGB : *Red Green Blue.*
TIFF: *Tag Image File Format.*
BMP : *Bitmap.*
JPEG: *Joint Photographic Experts Group.*
GIF: *Graphic Interchange Format.*
PNG: *Portable Network Graphics.*
SVG: *Scalable Vector Graphics.*
MP3: *Moving Picture Experts Group Layer-3 Audio.*
URL: *Uniform Resource Locator.*
ABox: Assertion component.
TBox: *Terminological component.*
FOIL: *First Order Inductive Learner.*
VOS : *Video Ontology System.*
VERL : *Video Event Représentation Langage.*
VEML: *Video Event Markup Language.*
ECCV : *European Conference on Computer Vision.*
PETS : *Performance Evaluation and Tracking System.*
TRECVID : *TREC Video Retrieval Evaluation.*
IA : *Intelligence Artificielle.*
XML: *eXtensible Markup Language.*
RDF: *Resource Description Framework.*
RDFS: *Resource Description Framework scheme.*
OWL: *Web Ontology Language.*
W3C: *World Wide Web Consortium.*
ROI: *Region of Interest.*
CAVIAR: *Context Aware Vision Using Image-Based Active Recognition.*
LSTM: *Long Short-Term Memory.*
CNN: *Convolutional Neural Networks.*
FHD: *Full High-Definition.*
UML: *Unified Modeling Language.*
OpenCV: *Open Computer Vision.*
CPU: *Central processing unit.*
GB: *GigaByte.*
RAM: *Random-Access Memory.*
SVM: *Support Vector Machine.*
KNN: *k-nearest neighbour.*
SED: *Surveillance Event Detection.*

Table des matières

Chapitre III : Solution Proposée basée sur l'approche ontologique

Annexe B : Cas pratique de mise en œuvre de notre système OVIS

Références

Liste des Figures

Liste des Tableaux

Introduction Générale

Contexte et motivation

Les avancées technologiques dans le domaine numérique ont engendré une quantité très importante de documents multimédia, et cela grâce à la diversité d'outils de capture tels que les téléphones portables, les caméras, les appareils photos, les tablettes, etc. Or cette masse très importante de données multimédia nécessite leur organisation, leur stockage dans des bases de données et leur réutilisation et partage pour des buts professionnels ou personnels à travers les réseaux sociaux ou sur Internet en général.

En effet, beaucoup de chercheur à travers le monde ont été motivés par le traitement d'images et vidéos qui représente un axe très important de recherches, connu sous le nom de vision par ordinateur (vision artificielle, vision numérique ou vision cognitive). Par ailleurs, les travaux de recherche dans ce domaine ont pour but d'analyser et de transformer la vision humaine sur un ordinateur et ainsi le doter d'une capacité qui se rapproche le plus à celle de l'être humain. Ainsi, l'objectif général se résume à la modélisation et la création de systèmes d'interprétation et de représentation d'images ou de scènes vidéo connues aussi sous le nom d'indexation et de recherche de vidéo.

Grâce à la masse importante de centaines de milliers d'heures de données vidéo numériques capturées, stockées et archivées, le problème d'indexation et de recherche vidéo a connu le jour. En outre, les avancées technologiques réalisées ces dernières années dans le domaine de l'informatique (espaces de stockage de plus en plus considérables, numérisation des données, etc.) ont permis d'augmenter l'utilisation de données vidéo par le grand public.

La richesse sémantique d'un document vidéo le rend important et plus expressif grâce à sa nature hétérogène. Néanmoins, sa structure reste plus ambiguë car il est difficile de mettre en évidence une structure unique dans un document vidéo intégrant simultanément de l'image, du texte et du son.

Outre les problèmes de stockage et d'archivage, d'autres problèmes d'utilisation, de recherche, de navigation et d'extraction se posent avec la croissance constante de masses de données vidéo. Par contre, la consultation des documents vidéo doit être en effet facile. Par conséquent, un processus de modélisation, d'indexation et de recherche doit être mis en place de manière à accélérer la recherche de l'information souhaitée.

Savoir manipuler de l'information vidéo nécessite un fort besoin dans diverses industries de production, d'archivage ou de distribution de contenu vidéo. De façon très informelle, les

systèmes manipulant la vidéo sont caractérisés par le traitement de données informatique exprimée dans divers média (son, image, texte). En effet, les éléments clés permettant cette intégration sont la puissance et les possibilités croissantes de la nouvelle génération d'ordinateurs personnels ou de serveurs de calcul et de stockage qui apparaissent sur le marché. Ces ordinateurs se distinguent de la génération précédente par leur capacité à manipuler l'ensemble de ces médias, leur stockage ainsi que leur échange sous une forme entièrement numérique. Il devient ainsi possible de restituer de l'information aussi bien sous forme audiovisuelle que textuelle et graphique à un ou plusieurs utilisateurs connectés à travers un réseau de communication.

Concernant la variété du contenu, chaque type de document vidéo possède sa propre structure qui le distingue des autres. En outre, ces documents vidéo peuvent être : des journaux télévisés, des émissions sportives, des films, des documentaires, des émissions de téléréalité ou de cuisine, des enregistrements vidéo de surveillance, etc. Dans notre travail de thèse, nous nous sommes intéressés à la modélisation, l'indexation et la recherche d'évènements vidéo dans un système de vidéosurveillance, comme l'illustre la **Figure 1**

Figure 1. Exemple d'un système de vidéosurveillance.

La motivation principale de notre recherche consiste à concevoir et réaliser un nouveau système d'indexation et de recherche permettant de faciliter l'accès à une base de données de vidéosurveillance, en tenant compte des captures vidéo prises par des caméras intérieures ou extérieures des lieux à surveiller.

Problématique de la thèse

La manipulation des données vidéo exige un besoin croissant des systèmes informatiques. En effet, ces systèmes doivent être capables de gérer ce type de données d'une manière efficace et précise. Dans un contexte réel, lorsqu'un utilisateur souhaite rechercher des vidéos, il est souvent plus pratique pour lui d'utiliser une information sémantique (un événement ou un concept spécifique) pour obtenir les réponses les plus pertinentes. Or les systèmes actuels ne satisfont pas vraiment ce besoin du fait que dans la plupart des cas les auteurs privilégient un type spécifique de média. En effet, beaucoup de systèmes d'indexation existants dans la littérature exploitent uniquement l'aspect visuel des vidéos, et ne considèrent pas leur aspect sémantique. À l'exception de quelques systèmes qui sont spécifiques aux applications multimédia comme la vidéosurveillance, les émissions sportives, etc., la plupart des systèmes d'indexation de l'état de l'art font le choix d'un processus automatique qui ignore le point de vue de l'utilisateur final.

La représentation d'un document vidéo diffère selon le domaine d'application (archivage, recherche d'information, édition, etc.). Une expression du contenu sémantique n'est pas utile à toutes les applications manipulant les vidéos. Ainsi une application d'archivage vidéo peut s'intéresser aux pixels de chaque image et aussi leurs positions spatiales pour des fins de compression, alors qu'une opération de montage vidéo manipule l'image entière comme unité de base et la considère comme l'unité élémentaire pour la vidéo.

La problématique de la représentation d'un document vidéo ou de son contenu aussi bien sémantique que structurel apparaît principalement lors des phases d'indexation, d'expression des requêtes et d'interaction avec l'utilisateur. D'une manière générale, on distingue deux intérêts essentiels pour représenter la vidéo :

- **La transmission ou le stockage :** on se place ici dans un contexte de codage d'information. Dans ce cas, il est nécessaire de passer par des étapes de compression qui produisent une représentation codée plus compacte et difficile à manipuler. D'un point de vue pratique, la phase de décompression servira à récupérer le maximum d'information originale d'une vidéo, avec le moins de dégradation possible selon une perception humaine. Plusieurs standards ont été alors définis, comme par exemple les normes MJPEG, MPEG, etc.
- **L'indexation et la recherche :** dans ce cas, il est plutôt question de manipuler la vidéo pour mettre en place une base d'indexation qui constitue une représentation virtuelle du document vidéo servant d'intermédiaire entre le document et les

besoins d'information exprimés sous forme de requêtes. Parmi les propositions qui ont été faites dans ce contexte, nous citons la norme MPEG7 et Dublin Core.

Pour assurer l'efficacité d'un processus d'indexation et de recherche, il faudrait tenir compte de la variation individuelle dans l'interprétation des documents vidéo. En raison de la nature visuelle du signal vidéo, les données vidéo sont perçues et interprétées différemment par des personnes différentes. Toutes les interprétations ne peuvent pas être représenter par des mots clés car il n'est pas possible de les prévoir toutes au moment de l'indexation. En outre, la représentation d'un petit segment de signal vidéo par un grand nombre de mots clés mènera à l'explosion de la base d'indexation. D'autre part, les mots clés ne peuvent pas représenter la nature temporelle des signaux vidéo ni les rapports sémantiques des descriptions du contenu (règles d'inférence et hiérarchie, etc.…).

Les documents vidéo ont un caractère multimédia qui fait que la recherche par le contenu présente un certain nombre de spécificités. Par exemple, un concept donné (personne, objet...) peut être présent sous différentes manières : il peut être visualisé, entendu, parler dans un discours, et la combinaison de ces cas peut également se produire. Naturellement, ces distinctions sont importantes pour l'utilisateur. Des requêtes impliquant le concept A comme par exemple : « rechercher les segments vidéo montrant une image de A » ou comme : « rechercher les segments vidéo dans lesquels on parle de A » sont susceptibles de produire des réponses tout à fait différentes. Dans le premier cas, on rechercherait A dans le flux visuel tandis que dans le second, on rechercherait dans le flux audio un segment dans lequel A est mentionné.

À l'heure actuelle, plusieurs propositions ont été faites pour l'indexation et la recherche par le contenu mais la plupart d'entre elles mettent en avant un cadre d'étude très restreint et spécifique ou bien proposent une approche basée sur un type spécifique de média (texte, image ou audio). Ceci ne donne pas toujours des résultats satisfaisants et efficaces. Il est donc impératif d'accorder plus d'attention aux aspects d'analyse et de modélisation du contenu vidéo d'une manière plus globale.

Généralement dans le domaine de la vidéosurveillance, on peut classer ces propositions en deux grandes catégories : la première est basée sur le bas niveau en utilisant des descripteurs de la vidéo et la deuxième est basée sur le haut niveau en utilisant l'aspect sémantique de la vidéo. Concernant la première catégorie, elle consiste à utiliser les données de bas niveau de la vidéo tels que les descripteurs de couleurs, de texture, de forme, etc. et dont l'inconvénient relève de sa précision qui est dans la plupart des cas très mauvaise et s'éloigne de la vision humaine [1, 2]. Par contre, la seconde catégorie se base essentiellement sur la sémantique de la vidéo et à

l'inverse de la précédente, fournit de bons résultats en termes de précision. Son inconvénient se résume dans la difficulté de la tâche à effectuer dans le cas où il s'agit d'une très grande base vidéo [3, 4]. Pour remédier à cet inconvénient, la combinaison de ces deux catégories devient alors une tâche essentielle pour garantir de bons résultats et indexer à la fois une grande base de vidéos. Cependant, cette combinaison génère un problème très important connu sous le nom de gap sémantique (voir Figure 2).

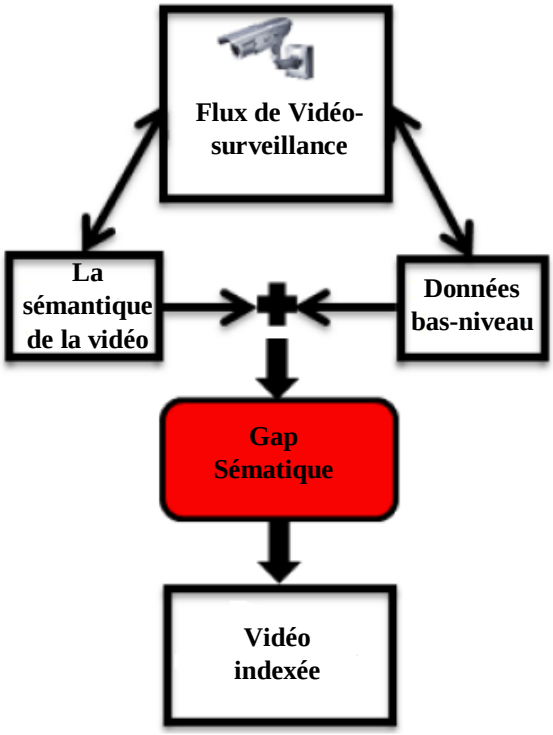


Figure 2. Le gap sémantique entre le haut niveau et le bas niveau d'une vidéo.

Généralement, le problème du gap sémantique caractérise la différence entre la sémantique d'une vidéo exprimée par des experts du domaine de la vidéosurveillance et l'interprétation des résultats obtenus à partir du module d'analyse de bas niveau. Afin de résoudre ce problème, plusieurs travaux de l'état de l'art se sont basés sur une approche ontologique mais sans aucune méthode définie [5]. Aussi, les cas des expérimentations réalisés sont limités aux évènements représentant un ou deux objets d'intérêt dans une séquence vidéo tel qu'un objet abandonné, une personne qui marche de gauche à droite, un avion qui décolle, etc. Malheureusement, les travaux de l'état de l'art n'ont pas traité les cas de plusieurs objets d'intérêt comme celui de la foule de gens (par exemple des gens qui se dispersent, des gens qui se regroupent, etc…).

Dans notre travail de thèse, on a conçu et implémenté une ontologie du domaine de la vidéosurveillance en tenant compte de tous les cas d'évènements et plus particulièrement le cas

de plusieurs objets d'intérêt apparaissant dans une scène vidéo. Ainsi, le problème du gap sémantique a été résolu par notre système d'indexation et de recherche vidéo OVIS [6].

Objectifs de la thèse

Le comportement des personnes appartenant à une foule de gens présents dans une scène d'une vidéosurveillance est analysé d'une manière différente. En effet, on s'intéresse au cas de la totalité de la foule afin d'en extraire les évènements pertinents d'une vidéo. Pour cela, une phase d'indexation est alors nécessaire afin d'effectuer une recherche ultérieure. Notre travail de thèse concerne l'indexation et la recherche de tous les évènements issus d'un flux de vidéosurveillance en tenant compte de plusieurs objets d'intérêt.

Ainsi, les objectifs de notre thèse sont doubles : le premier consiste à proposer une modélisation des objets, actions et évènements d'une vidéo à l'aide d'une ontologie de vidéosurveillance et le second propose un nouveau système d'indexation et de recherche vidéo appelé OVIS utilisant les règles d'inférences SWRL [7].

Contributions de la thèse

Les contributions essentielles de notre thèse peuvent être résumées comme suit :

- *Une étude détaillée de l'état de l'art des approches et des systèmes d'indexation et de recherche vidéo* :
  - Identification des inconvénients et limites des approches existantes.
  - Etablir un tableau comparatif de notre ontologie de vidéosurveillance avec celles existantes dans la littérature selon les métriques suivantes : la consistance, le formalisme, la largeur du domaine de représentation et la conceptualisation.
- *Conception et implémentation d'une ontologie de vidéosurveillance* :
  - L'ontologie proposée respecte une convention de nommage syntaxique bien définie afin d'assurer sa consistante et sa complétude.
  - Elle décrit tous les objets, les actions et les évènements qui peuvent apparaitre dans une vidéosurveillance, que ce soit dans le domaine académique ou industriel.
  - L'avantage de notre ontologie de vidéosurveillance est qu'elle est non seulement d'usage multiple puisqu'elle peut être utilisée dans l'indexation et la recherche de vidéos, mais aussi dans la création des benchmarks, dans la représentions des scènes ou dans la représentation des vidéos en général.

- *Indexation des évènements de foule de gens* : Afin de prouver l'efficacité de notre approche ontologique, nous avons réalisé un système d'indexation et de recherche vidéo OVIS utilisant les règles d'inférences SWRL qui sont réparties en quatre catégories :
  - La catégorie de règles SWRL de distance qui regroupe les différents Bounding box détectés par le module de bas niveau de chaque frame d'image appartenant à une séquence vidéo en un seul Bounding box.
  - La catégorie de règles SWRL de suivi qui permet d'identifier les différents groupes de personnes occurrents dans une séquence vidéo.
  - La catégorie de règles SWRL d'évènements qui analyse le comportement de chaque groupe d'objets détecté pour extraire l'évènement voulu.
  - La catégorie de règles des relations d'Allen qui illustre les différentes relations d'Allen existantes entre des scènes vidéo.

Nous allons étudier plus en détails ces différentes contributions dans le chapitre 4, qui concerne l'implémentation, l'expérimentation et la comparaison de notre système d'indexation et de recherche vidéo OVIS.

Structuration de la thèse

Notre thèse est structurée comme suit :

Le chapitre 1 concerne les généralités sur les documents vidéo.

Le chapitre 2 explique le domaine d'indexation et de recherche vidéo et fournit un état de l'art détaillé des systèmes d'indexation et de recherche vidéo. Il détaille aussi notre étude comparative ainsi que toutes nos contributions scientifiques.

Le chapitre 3 est consacré à une introduction au domaine des ontologies avec la présentation de la solution proposée à savoir notre ontologie de vidéosurveillance et ses différentes applications académique et industrielles.

Le chapitre 4 détaille l'architecture générale de notre système d'indexation et de recherche vidéo OVIS avec une modélisation UML de notre prototype. Il fournit aussi les résultats des expérimentations réalisés en utilisant les benchmarks PETS 2012 et TRECVID 2016 avec les différentes évaluations en utilisant :

- Les métriques telles que : la précision, le rappel, etc…
- Le temps de détection des évènements par rapport à la scène vidéo, et

- Les relations d'algèbre d'Allen.

Enfin, nous conclurons notre manuscrit de thèse en résumant nos contributions scientifiques et en présentant les perspectives futures.

Chapitre I : Généralités sur les Documents Vidéo

1.1. Introduction

De nos jours, plusieurs types de documents multimédias existent sous différentes formes et types d'extensions. Parmi cette variété de modèles, on peut trouver les documents textuels, sonores, graphiques et audiovisuels. Dans notre travail de thèse, on s'intéressera aux documents audiovisuels. En effet, avec les nouvelles technologies de stockage vidéo, on peut trouver des documents vidéo qui génèrent à la fois : l'animation, le son et même le texte lors des présentations des médias. En outre, ces documents vidéo représentent plusieurs domaines tels que : les journaux télévisés, les vidéos du web, le sport, les divertissements et documentaires, la vidéo surveillance etc….

Dans ce chapitre, nous allons détailler les différents types de média à savoir le texte, l'image, l'audio ou la vidéo. Nous allons nous focaliser par la suite sur les documents vidéo qui représentent l'intérêt de notre problématique de thèse en expliquant leurs différentes caractéristiques. Pour cela, nous évoquerons la modélisation d'un document vidéo en citant les différentes approches à savoir l'hiérarchique, en strate et par objet. Enfin, nous terminerons par une conclusion.

1.2. Le médium ou média

La notion de medium signifie originalement en latin "*milieu, centre*" mais aussi "*lieu accessible à tous, à la disposition de tous, exposé aux regards de tous*". Le mot prend plus tard le sens d'intermédiaire et de moyen de communication de la pensée [8]. Le mot média est en fait un raccourci du mot anglophone mass media, ce dernier englobe l'intégralité des moyens de communication par voie écrite, radiophonique et télévisée. La signification du mot média est universalisée dans le monde de l'informatique en considérant qu'un média est semblable à un moyen de transmission, de stockage, ou de présentation d'informations.

Par multimédia, on entend un produit ou un service qui associe des informations d'origines diverses (texte, image, son, vidéo, etc.) en offrant à l'utilisateur la possibilité de les consulter de façon interactive, le tout étant disponible sous forme numérique sur divers types de support. Le mot multimédia est souvent utilisé à propos des techniques qui permettent d'associer des informations d'origines diverses tels que le texte, le son, des images fixes ou animées, de la vidéo, etc. En réalité, le multimédia est bien plus qu'un ensemble de techniques. C'est un moyen de communication et de diffusion d'informations à part entière tout comme la presse, le cinéma,

la radio ou la télévision.

Pour qu'on puisse parler de multimédia, il faudrait considérer les quatre caractéristiques suivantes :

- L'existence et la mise à la disposition de l'utilisateur des éléments informationnels d'origine et de nature diverses tels que le texte, l'image fixe ou animée, le son ou la vidéo.
- L'organisation structurée de ces divers éléments suivant un scénario basé sur des liens logiques permettant à l'utilisateur de se déplacer ou de dialoguer à son propre gré de manière interactive.
- Le lien de tous les éléments sous forme d'un ensemble cohérent formant un produit ou service original. On peut ainsi parler d'une œuvre multimédia tout comme on parle d'une œuvre cinématographique.
- La mise à la disposition du produit ou service à l'aide d'un support faisant appel aux techniques de numérisation et de compression.

Généralement, on classifie le produit ou service multimédia en deux types suivant le moyen d'accès de l'utilisateur, "en ligne" ou "hors ligne".

- **En ligne (online) :** l'utilisateur accède au produit ou service à l'aide d'un réseau. Le produit ou service multimédia est alors stocké à distance sur une machine autre que le poste de travail de l'utilisateur (ordinateur, console, GSM, PDA, télévision ou autre). Cependant, le réseau servant à diffuser les informations est en général Internet et il arrive même souvent que l'on dise Internet à la place de multimédia en ligne. Par ailleurs, le marché du multimédia en ligne est très ouvert et en pleine croissance.
- **Hors ligne (offline) :** l'utilisateur accède au produit ou service multimédia à partir de sa machine ou poste de travail sans être connecté à un réseau. Le marché du multimédia hors ligne est plus restreint et souvent plus ciblé sur certains secteurs comme par exemple celui des jeux vidéo.

Une application ou "œuvre" multimédia peut fonctionner de plus en plus fréquemment en utilisant simultanément ou alternativement les deux types online ou offline. C'est le cas d'une application éducative sur CD-ROM ou DVD qui contient des bonus accessibles sur Internet, ou encore d'un jeu vidéo sur console pouvant également fonctionner sur plusieurs consoles en réseau.

Chaque média est l'objet d'études particulières visant à faire avancer sa compréhension et sa maîtrise. La représentation des données mono média dans un format informatique compréhensible par une machine, est appelée codage. Des efforts considérables sont faits afin de définir des codages adéquats pour chaque média afin de faciliter certaines opérations [9].

1.2.1. Le média texte

Le média texte est un média factice. C'est le plus ancien au monde et il a fait l'objet des premiers développements en informatique. Ses informations sont conceptuellement bien intégrées dans nos modèles courants. Ce qui les rend simples à modéliser au sein des systèmes informatiques. Le texte fondamental est une séquence de caractères d'un alphabet. Il est traditionnellement découpé en mots, phrases, paragraphes, sections et chapitres au sein d'une œuvre littéraire comme un article, un livre ou un texte enrichi. Il ajoute beaucoup d'informations de mise en forme en adhérant à chaque caractère des paramètres de présentation, comme la police de caractère (groupe de caractères, de caractéristiques, de formes identiques), la casse, le gras, l'italique, le soulignement ou bien encore la couleur. Lorsque le texte est mis en page, chaque caractère possède un alignement relatif aux caractéristiques de la page (tailles, marges, etc.).

Bien que moins dense que les autres médias, le texte reste encore aujourd'hui l'élément privilégié pour exprimer le sens et expliciter la signification des autres médias grâce à son intuition et sa simplicité [10].

Les formats texte sont : Le format DOC (Microsoft Word) utilisé pour coder les documents engendrés par le traitement de texte Microsoft Word, Le format ODT (Open Document Text) décrivent le contenu et les styles utilisés dans le document, regroupés dans une archive ZIP, et le format PDF (Portable Document Format) est un format ouvert mais pas compréhensible directement. Il permet d'intégrer texte, images matricielles et/ou vectorielles...etc.

1.2.2. Le média image

Une image est une représentation visuelle, voire mentale, de quelque chose (objet, être vivant et/ou concept) Elle peut être naturelle (ombre, reflet) ou artificielle (peinture, photographie), visuelle ou non, tangible ou conceptuelle (métaphore), comme elle peut entretenir un rapport de ressemblance directe avec son modèle ou au contraire y être liée par un rapport plus symbolique. Concernant la sémiologie ou sémiotique, l'image a développé

tout un secteur de sémiotique visuelle et elle est conçue comme produite par un langage spécifique.

Une des plus anciennes définitions de l'image est celle donnée par Platon : « *J'appelle image d'abord les ombres ensuite les reflets qu'on voit dans les eaux, ou à la surface des corps opaques, polis et brillants et toutes les représentations de ce genre* ». Le mot image en français vient du latin imago, qui désignait autrefois les masques mortuaires.

Les images sont des constituants bidimensionnels composés de points, ou de pixels. Chacun d'eux possède une couleur ou hypothétiquement une transparence. Les couleurs contiennent des codages propres tels que le triplet RGB qui est de couleur Bleu, Vert et Rouge. Cette définition nous permet de voir la représentation du médium par le système informatique. Par contre les courbes et les formes sont perçues par l'œil humain comme un rassemblement de pixels de couleurs et de positions proches, que lui seul peut interpréter. Le domaine de la génération d'image, ou infographie, s'est énormément accru et a conçu des outils performent comme Adobe Photoshop.

Beaucoup de formats binaires d'image ont été normalisés, comme TIFF, BMP, JPEG, GIF ou PNG. Ils acquiescent à conserver des informations sur les pixels de l'image et les échanger entre les systèmes informatiques, chacun facilitant certaines opérations comme l'affichage progressif ou bien la compression. Suivant la norme XML, des formats textuels d'image sont apparus, comme le standard SVG (Scalable Vector Graphics). Les formats d'image continuent d'évoluer en intégrant des techniques issues d'autres domaines. Par exemple les smart graphiques [11], étant aptes à s'acclimater à des conditions distinctes, comme les tâches pour lesquelles elles seront utilisées ou les paramètres réseaux de transmission des images. La représentation de bas niveau d'une image ne peut avoir que deux dimensions. L'image peut aussi être une superposition de plusieurs images simples, tout en réglant l'alignement et la transparence pour clarifier leurs positions relatives.

1.2.3. Le média audio

Le son ou audio est l'un des principaux médias utilisés aujourd'hui avec la vidéo. C'est un média temporel dont l'information est représentée par un signal périodique et continu, qui est produit par des périphériques comme les haut-parleurs et perçu par l'oreille de l'utilisateur humain. Le son étant une donnée temporelle qui peut évoluer vite, il requiert une grande

quantité d'informations [12].

Il existe des méthodes de compression qui sont générales pour les représentations continues. Ces dernières sont fournies par les formats sonores. Ces méthodes de compression offrent déjà un gain pour les signaux physiques, ou spécifiques dans le cas direct, il existe un modèle psycho acoustique format MPEG-1, Layer 3 ou aussi MP3. Des algorithmes de compression sont définis par leurs implémentations qui sont aussi des composants logiciels appelés Codecs. En outre, le codage relatif des informations caractérise les valeurs en termes d'écart par rapport aux précédentes valeurs et accorde un gain additionnel. Les codages facilitent plus ou moins certains traitements d'informations parmi lesquels on peut trouver la décompression ou l'accès non séquentiel. Par ailleurs le son peut instaurer une dimension annexée dans certains contextes, comme on peut le voir nettement dans certains jeux vidéo.

1.2.4. Le média vidéo

Le médium vidéo est représenté comme une succession d'images animées et d'audio synchronisés dans un flux de données (Stream) [13]. La diffusion de la vidéo sur des ordinateurs s'est accompagnée du développement d'outils de compression permettant de stocker de gros volumes de données. Il s'agit du médium le plus gourmand en mémoire et par conséquent il requiert encore plus de compression que les autres médias. Les opérations liées à la vidéo concernent le stockage, la recherche, la synchronisation, l'édition, la gestion des effets spéciaux, la conversion entre formats, etc.

a. La forme

D'un point de vue physique (ou informatique), un document vidéo est un édifice de sous-médias ou « pistes » agencés suivant un axe temporel. Chaque piste est présente sous l'allure d'un flux d'éléments et les flux correspondants aux différentes pistes sont synchronisés entre eux. Ces différents flux peuvent comporter des images, du son ou du texte :

- **Son**

La majorité des documents vidéo comprennent aussi une (ou plusieurs) piste(s) « audio ». Les acteurs de cette piste sont des exemples émis à une fréquence ancrée (typiquement de 16000 à 48000 par secondes). Une piste audio peut encore être structurée de plusieurs flux de tels éléments en parallèle (cas de la stéréo, du codage à 6 canaux). Un document (ou flux) vidéo peut comporter plusieurs pistes audios en parallèle (correspondant à plusieurs langues par

exemple).

- **Texte**

D'autres documents vidéo incluent aussi une (ou plusieurs) pistes textuelles. Les éléments de cette piste ne sont communément pas formulés à une fréquence fixe mais précisément par des groupes accompagnés d'informations autorisant de les synchroniser avec les autres flux (temps de début et de fin ; dans le cas des pistes image et son, la synchronisation se fait sur la base de l'émission régulière et à une fréquence fixe des éléments).

b. Le contenu

Les contenus des documents vidéo sont extrêmement variés (journaux télévisés, vidéosurveillance, films, documentaires, publicité, etc.…). La plupart du temps, ces documents ont une (ou plusieurs) structures internes. Semblable aux documents eux-mêmes, ces structures peuvent être très diversifiées. Ces structures apparaissent souvent de manière bien arrangée : un document est décomposé hiérarchiquement en unités plus petites selon un arbre (pas forcément régulier, et notamment parfois de profondeur variable). Une désintégration classique (mais particulière) illustre par exemple les niveaux « vidéo » (document dans son ensemble : suite de séquences), « séquence » (suite de scènes), « scène » (suite de plans), « plans » (suite d'images), « images » (suite de régions), « régions » (ensemble de pixels).

Pour des raisons adaptées et de sens, toutes les unités appartenant à un segment continu correspondent à un intervalle temporel dans lequel on considère toujours les différentes pistes ensemble et alignées. Toujours pour des raisons pratiques, les éléments inférieurs de la hiérarchie ne se recouvrent pas entre eux. Ils se suivent temporellement les unes les autres et, ensemble, ils recouvrent l'élément de niveau juste supérieur dans l'arbre hiérarchique.

Une séquence vidéo est une indexation d'éléments d'un document numérique vidéo selon un organisme de règles. Il est question d'images numériques disposées chronologiquement avec éventuellement une bande de son. Un document numérique vidéo peut être estimé comme une seule ou un ensemble de séquences vidéo. Une séquence vidéo est identifiée par son instant de départ et de fin sur l'échelle de temps du document vidéo numérique. En général les documents vidéo sont le résultat d'un montage qui consiste à coller l'un après l'autre des plans avec d'éventuels effets de transition.

c. Définition des documents vidéo

Par définition, Un flux est un déplacement d'éléments dans le temps et dans l'espace. Dans le cas de document vidéo, ces éléments sont les données audio et visuelles. Un document vidéo (ou document audiovisuel) peut être perçu comme une superposition de flux mais la manière dont il a été composé et la mise des éléments ensemble ne sauraient se réduire à une superposition de flux. Un document audiovisuel possède une structure physique décrivant la mise des images ensemble et la synchronisation du son plus ou moins complexe, tandis que sa structure logique, non explicite, recouvre toutes les analyses possibles du document. Dans ce dernier cas, il s'agit de la notion du contenu de document, qui désigne l'ensemble des informations formant une vidéo (image, audio, texte).

- **Plan vidéo**

Les plans s'accordent à leur tour à des unités sémantiquement plus adhérentes appelées scènes. Dans le langage cinématographique, le découpage d'une vidéo désigne la division en plans et scènes. Une scène est constamment vue comme étant l'unité de référence de la vidéo. C'est à dire, quand on montre une vidéo à quelqu'un qui ne l'a pas vu, on définit un thème général de chaque scène. D'un point de vue interprétation, rien ne prouve qu'une unité sémantique de la vidéo coïncide nécessairement avec la division en plans. La fin d'une unité sémantique peut ne pas se situer sur une transition de plan. Toujours est-il que les transitions dans le contenu audio paraissent plus proches des changements de thème dans le document vidéo.

Un plan est défini dans un cadre de montage vidéo depuis une série d'images établie par une seule caméra. La segmentation de la vidéo en plan peut être faite par un processus automatique qui se base sur la détection de transitions entre les plans. En contrepartie, pour segmenter une vidéo en parties selon une description sémantique : unité de lieu (scène) ou unité de sujet (séquences), on doit communément faire appel à un opérateur humain.

- **Scène**

Une scène est composée d'un ensemble de plans ayant une même unité de lieu. Au niveau visuel, une scène vidéo entraîne des problèmes tels que par exemple :

- Comment enclaver une scène ?

- Sur quelle logique doit-on se reposer pour caractériser l'enchaînement des scènes ?

➢ **Séquence**

Une séquence centralise différents plans et scènes comme c'est illustré dans la Figure 3. Elle constitue une unité de sujet (par exemple un reportage dans un journal télévisé).
Au sein de ces trois unités, le plan représente l'entité de base. Au niveau syntaxique, une telle structuration ressemble à celle d'un document textuel (mot, phrase, paragraphe).

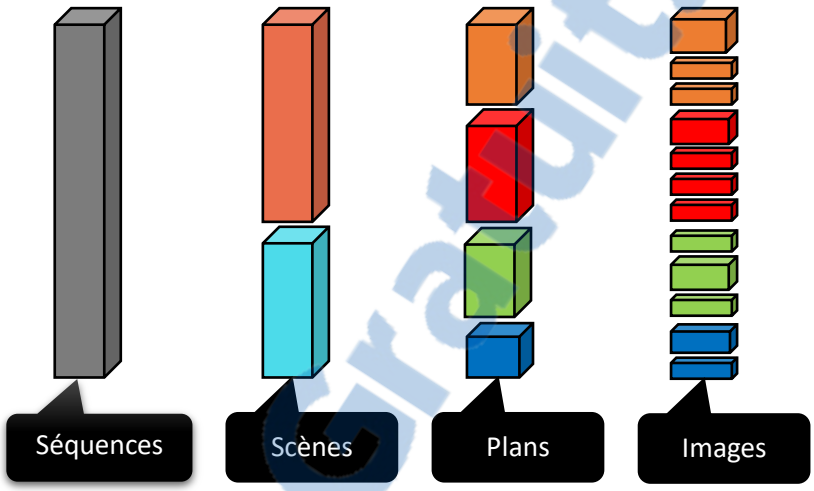


Figure 3. La segmentation temporelle d'une séquence vidéo.

➢ **Unité audiovisuelle (UAV)**

Par unité audiovisuelle, on définit une entité abstraite représentant un segment quelconque de document vidéo. Une unité audiovisuelle se rattache au type de l'information ainsi que la manière selon laquelle ce type d'information est segmenté. Une unité audiovisuelle sera immatriculée par la manière dont le document vidéo est segmenté. Par exemple, si on prend le cas d'une suite d'images on peut considérer qu'un plan représente une unité audiovisuelle. Par contre, pour le contenu audio, la segmentation selon le changement de locuteur peut constituer l'unité de repérage.

➢ **Eléments d'intérêts**

On entend par élément d'intérêt un concept (visuel ou audio) décrivant de manière pertinente l'unité audiovisuelle. Il se manifeste donc par n'importe quel élément au moment où il a été repéré dans le document comme élément d'intérêt. Donc il y aura un certain nombre d'éléments d'intérêt de façon à mener une analyse particulière sur le document vidéo.

- **Les métadonnées (ou méta-information)**

Ceux des données structurées qui décrivent d'autres renseignements. Dans le cas de la vidéo, les métadonnées détaillent les informations ressources sur le document. Ces ressources peuvent être affiliées directement dans la description.

Les métadonnées peuvent être indiquées comme étant des données relatives à d'autres données (données sur des données). Par conséquent, une notice sur le document peut être considérée comme métadonnée.

Utilisées dans le contexte de la recherche d'information, les métadonnées sont perçues comme des informations de fond qui analysent le contenu et autres propriétés et caractéristiques des données. On démêle trois types de catégories :

- **Métadonnées techniques :** elles donnent les informations techniques sur le programme (exemple : format d'image en TV, format d'enregistrement, type de support).
- **Métadonnées administratives :** elles donnent les renseignements fondamentaux pour la réalisation de tâches administratives liées à l'exploitation des contenus (droits d'auteur, droits de diffusion…).
- **Métadonnées descriptives :** elles caractérisent l'étendue des documents. Elles peuvent être relatives à un document accompli ou à des séquences de document. Dans certains cas, il peut y avoir une description scène par scène avec le nom des intervenants ou acteurs, le dialogue, le résumé de la scène, des thèmes actionnaires.

- **Image-clé (key frame)**

Une image-clé est une image qui assagit toutes les informations inéluctables à son affichage. C'est une image complète qui va servir de référence pour la reconstruction des images partielles de la séquence. Ainsi, une image-clé est souvent celle qui correspond à l'image la plus similaire dans le plan.

- **Les mouvements de caméra**

Il est très rare qu'un film soit composé uniquement d'une succession de plans fixes. Dans la très grande majorité des cas, la caméra est mobile et permet ainsi d'accompagner l'action. Ce mouvement de la caméra peut être discret ou captif [14]. Lorsqu'il est discret, il concède de suivre l'action, un personnage, tout en sachant se faire oublier. C'est le cas d'un panoramique ou d'un mouvement lent et fluide que l'on ne remarque pas, à condition qu'il ne soit ni en avance, ni en retard sur l'action. Cette dernière condition résume tout l'art d'un mouvement de

caméra réussi. Sauf si la caméra est en avance ou en retard sur l'action, son mouvement devient immanquablement captif. Il est discerné comme tel par le spectateur et trahit en quelque sorte la présence du réalisateur. Ce genre d'effet peut être voulu et permet de souligner une action particulière. La description des mouvements de caméra est réalisée par une combinaison de mouvements élémentaires. De plus, le zoom, qui n'est pas à proprement parler un mouvement de la caméra, est un degré de liberté supplémentaire qui leur est souvent associé.

d. Description du document vidéo

- **Description de bas niveau**

Les caractéristiques de bas niveau d'une vidéo sont des descripteurs physiques qui peuvent être extraits automatiquement par des algorithmes et sans connaissance particulière du contexte de la vidéo. Ces données correspondent généralement à des interprétations en termes de couleur, de forme, de structure ou de déplacement d'objets. Ces informations résultent de l'analyse de chaque image ou de segments d'images de la vidéo. Le plus souvent, les techniques d'extraction d'information liées à l'analyse d'une seule image, visent à la segmenter en régions afin d'en extraire des informations concernant les couleurs, textures et formes qu'elle contient. C'est ce qui est proposé notamment par les systèmes d'interrogation de bases d'images QBIC [15].

- **Description structurelle**

Les documents vidéo hiérarchiquement structurés en séquences, scènes, plans et images sont maintenant largement acceptés. Une structure semblable reflète le processus de création de la vidéo, les différentes étapes du montage. Les plans sont définis comme des séquences continues d'images prises sans arrêter la caméra. Les scènes sont définies comme des suites de plans contigus qui sont sémantiquement reliés.

- **Description de haut niveau**

La description sémantique du contenu d'une vidéo s'appuie sur la notion d'annotation. Une annotation représente une description symbolique de la vidéo ou d'un segment particulier de la vidéo. Bien que dans certains domaines ciblés tels que le sport [16] ou alors les journaux d'informations [17], il soit possible de réaliser une extraction automatique des annotations, la définition des annotations est réalisée manuellement par l'utilisateur, avec l'aide d'un logiciel d'annotation [18].

Les annotations sont des textes qui expliquent le contenu d'une séquence vidéo (document entier, scène, plan, image ou même objet vidéo). Pour cela, Il existe deux aspects distincts qui peuvent être exprimés par des annotations [19].

- La description des séquences vidéo qui présente les personnes, les choses, les scènes ou les événements qui apparaissent dans la séquence vidéo.
- L'interprétation des séquences vidéo qui énonce la teneur de la séquence vidéo ou des objets qui la structurent.

Les attributs sémantiques de la vidéo, qu'on appelle également caractéristiques de haut niveau, sont des informations qui donnent la description de cette vidéo selon l'adaptation humaine. L'interprétation des séquences vidéo dépend de deux éléments qui peuvent être définis comme suit :

- Le niveau de connaissance et la manière selon laquelle l'interprétateur la perçoit détermine son objectif lors de l'observation. En effet, les caractéristiques sémantiques sont étroitement liées au domaine d'utilisation et à la compréhension de l'interprétateur ; ce qui donne une certaine subjectivité aux caractéristiques.
- Le texte, contenant les caractéristiques sémantiques, peut être issu de plusieurs sources : texte entourant les vidéos dans le cas où les vidéos se trouvent à l'intérieur d'un autre document (ex. une vidéo se trouvant sur une page Web). On exploite alors le texte qui entoure la vidéo pour l'annoter. Ce texte peut être un ensemble de mots décrivant la vidéo, le titre de la vidéo, le nom de l'auteur, l'URL (Uniform Ressource Locator) de la page Web, etc.

Les avantages des méthodes de représentation des séquences vidéo basées sur des annotations sont les suivants :

- Le processus d'interrogation est classique, il s'adresse à des données alphanumériques.
- Les résultats sont plus absolus que dans le cas des méthodes basées sur le contenu des séquences vidéo.

Quant aux inconvénients, ils se manifestent comme suit :

- Le coût très élevé du processus d'annotation.
- La variabilité des annotations en fonction de la subjectivité de la personne qui a réalisé l'annotation.

e. Modélisation d'un document vidéo

On trouve dans la littérature trois approches de modélisation du contenu d'un document vidéo :

- **La modélisation hiérarchique**

Elle permet d'associer une description sous forme d'arborescence du contenu du document. Cette forme de modélisation est souvent liée à la segmentation temporelle. Une modélisation hiérarchique donne naissance à une représentation de la vidéo selon une structure arborescente dont la racine est le document vidéo. Une première phase consiste à découper récursivement le document vidéo en des unités plus petites, en général jusqu'au niveau du plan. Une deuxième phase consiste à choisir dans chaque plan une ou plusieurs image(s) clé(s). Ces images étant ensuite décomposées en régions ou en objets visuels (voir figure 4).

La modélisation hiérarchique des documents vidéo se base généralement sur le plan comme unité élémentaire et sur l'organisation de ces plans en unités de plus haut niveau sémantique, telles que les scènes. Celles-ci peuvent être alors regroupées au sein du document. Dans d'autres unités appelées séquences, la structure du document est alors une structure arborescente similaire à une structure documentaire textuelle classique (chapitre, section et paragraphe). La structure arborescente est souvent conçue suivant une hiérarchie document/séquence/scène/plan/image-clé/régions mais d'autres types de hiérarchies sont également possibles.

Des caractéristiques de niveau sémantique (classes, événements) ou de niveau signal (descripteurs) peuvent être associées aux différents niveaux de l'arborescence. Elles proposent une structure hiérarchique pour la représentation des documents vidéo par le contenu. Cette structure se base sur une modélisation sous forme de graphes et prend en compte les descriptions temporelles dans le document [20]. Un certain nombre de caractéristiques s'attachent aux différents segments de leur structure hiérarchique filmique : les caractéristiques générales (auteur, date, titre) sont adhérées au document dans son ensemble et des descriptions textuelles aux scènes. Les plans sont, quant à eux, caractérisés par des techniques primitives (mouvement de caméra, angle, profondeur de champ).

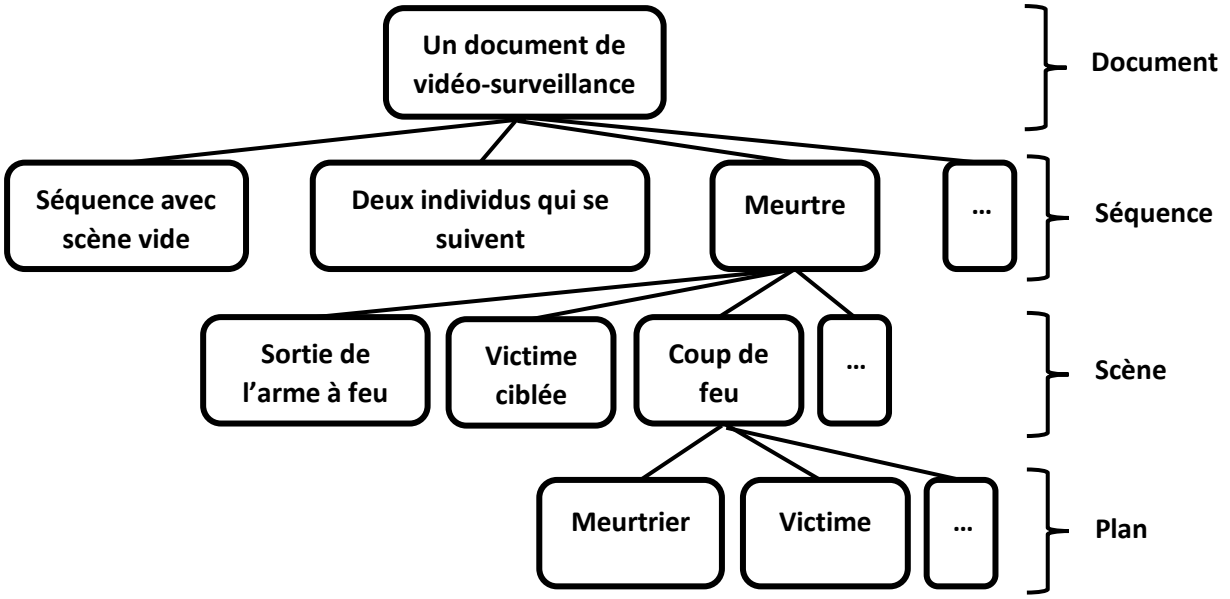


Figure 4. Modélisation hiérarchique d'une vidéo.

- **La modélisation en strates ou « stratification »**

Elle associe des annotations aux documents ou aux parties du document vidéo. La stratification est indépendante de la segmentation temporelle du document. Une strate est une liste de segments de vidéos auxquels est attachée une annotation. Un segment est un intervalle d'images fixes contiguës. Une strate regroupe des segments d'images qui partagent une sémantique commune, représentée par l'annotation. Chaque strate est associée à une liste de segments de vidéos ordonnés chronologiquement. Comme le montre la Figure 5, les strates d'une vidéo peuvent avoir des segments en commun. Cela veut dire que les objets, évènements ou actions étouffés dans les annotations respectives de ces deux strates se dévoilent ou se produisent simultanément dans les segments d'images qui se chevauchent. Pour conclure, la puissance d'expression d'un modèle d'annotation est cependant liée à sa capacité à définir finement des strates, les conditionnelles relations (ensemblistes, temporelles…) entre les strates, et les liens entre strates et annotations associées. Non moins important est le choix d'un formalisme de représentation de connaissances (logique, relationnel, objets, graphes conceptuels, réseaux sémantiques…) pour représenter les annotations chargées de la description sémantique de haut niveau.

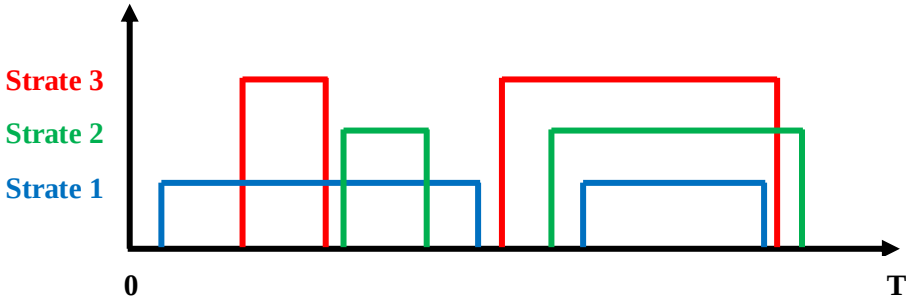


Figure 5. Stratification d'une vidéo

- **La modélisation par objet**

Elle nous permet de détailler le contenu avec des objets audiovisuels. Elle prend souvent la même structure que la modélisation hiérarchique. Nous apercevons que le rôle de ces différentes approches de modélisation des documents (en strates, hiérarchique ou objets) est de permettre de structurer le contenu d'un document vidéo pour éclairer son utilisation dans des applications multiples telles que la recherche d'information par exemple. À l'aide d'objets audiovisuels la description du contenu dans un document vidéo se fait (voir Figure 6) et chaque objet audiovisuel est évoqué par une description symbolique et des descripteurs (indices). La modélisation à base d'objets vise surtout à décrire le contenu sémantique du document vidéo. Par exemple, le contenu de la piste image n'est plus considéré comme un tableau de pixels (analyse bas niveau) mais comme un organisme d'objets (analyse conceptuelle) ayant chacun ses propres caractéristiques visuelles. Cette description a spécifiquement pour objectif une meilleure structuration de la représentation des vidéos pour ainsi permettre par exemple une navigation à différents niveaux de granularité (la vidéo, les scènes, les plans, les objets, etc.). La modélisation du contenu basée sur les objets est utilisée pour la manipulation du contenu et aussi pour la segmentation spatio-temporelle du document vidéo.

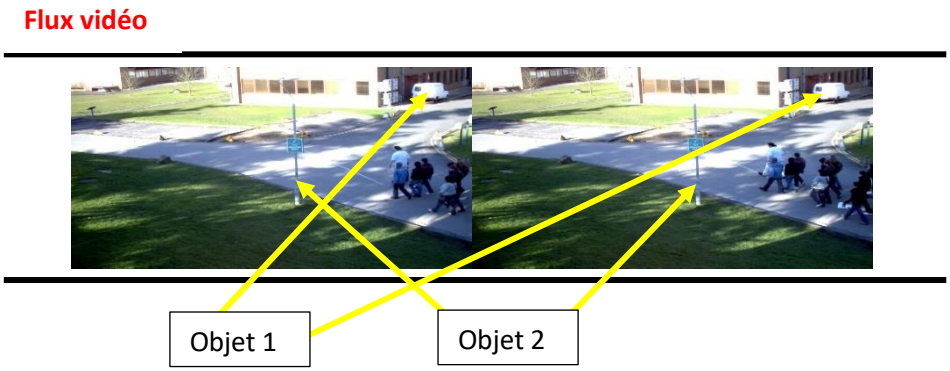


Figure 6. Le concept d'objet.

1.3. Conclusion

Dans ce chapitre, on a vu les différents types de média à savoir le texte, l'image, l'audio et la vidéo. Nous avons insisté sur les documents vidéo qui sont au cœur de notre problématique de thèse. Ces derniers représentent une mixture d'images animées synchronisées avec de l'audio et d'un commentaire sous forme de texte descriptif. Nous avons défini aussi une séquence vidéo comme une succession de scènes composés de plusieurs plans à la fois. Ces

différents plans sont composés à leur tour de plusieurs images qui représentent la plus petite unité de structuration d'une vidéo. Nous avons aussi présenté dans ce chapitre les différentes approches de modélisation des documents vidéo en utilisant les modèles : hiérarchique, en strate ou le plus utilisé à objets.

Dans le chapitre 2 suivant, nous allons détailler l'état de l'art des différentes approches et systèmes d'indexation et de recherche des vidéos.

Chapitre II : Indexation et Recherche des Vidéos (Etat de l'Art)

2.1. Introduction

Jusqu'à ce jour, un très grand nombre de chercheurs s'intéresse à l'étude du comportement humain et au suivi des objets dans une application de vidéosurveillance. Cette dernière nécessite alors des outils qui permettent d'organiser les vidéos pour faciliter l'accès et de rendre leur localisation plus rapide et plus efficace. Cela conduit aussi à l'exigence d'une gestion efficace des données vidéo qui ouvre la voie à de nouveaux domaines de recherche, tels que l'indexation et la recherche de vidéos qui respectent leur contenu spatio-temporel, visuels et sémantiques.

Les systèmes d'indexation conçus ne sont pas encore assez performants car ils exploitent le texte et le contenu. Cela donne la possibilité d'effectuer des recherches qui peuvent être basées sur le texte rattaché à la vidéo (*contenu sémantique*) ou basées sur le contenu visuel en utilisant ce qu'on appelle les caractéristiques de "bas niveau" (*couleur, forme, texture*). Par conséquent, ce processus d'indexation génère un problème majeur connu sous le nom de fossé sémantique. Ce dernier est défini comme une divergence entre la représentation de bas niveau d'une vidéo et son interprétation sémantique par les experts du domaine.

Dans la plupart des travaux de recherche, le mouvement humain est considéré comme une information de bas niveau qu'il faudra prendre en charge. Pour cela, la plupart des méthodes d'indexation et de recherche des vidéos utilisent les blobs pour l'extraction d'information de bas niveau et les ontologies pour l'extraction d'information de haut niveau. En effet, le problème d'interprétation de vidéo se pose toujours, par la présence au fossé sémantique existant entre le bas niveau et le haut niveau d'une vidéo.

Ce fossé sémantique représente alors la tâche la plus dure à résoudre dans le processus d'indexation et de recherche vidéo. En effet, l'interprétation d'une donnée brute exprimée en bas niveau sous forme d'une couleur ou d'une texture doit traduire réellement une donnée représentée en haut niveau sous forme de concept (objet, personne, animal, etc.) et qu'on doit utiliser dans le processus final d'indexation et de recherche vidéo.

Pour cette raison, plusieurs méthodes d'indexation et de recherche des vidéos existent dans la littérature et se basent sur des méthodes d'annotation manuelles, semi-automatiques ou automatiques des vidéos. L'annotation semi-automatique est la plus efficace et la plus utilisée dans les systèmes d'indexation et dont la recherche de vidéo utilise des mots clé (indexes) ou des vidéos similaires.

Dans ce chapitre, nous allons présenter le problème d'indexation et de recherche des vidéos

et ses caractéristiques. Un état de l'art sera présenté pour décrire toutes les méthodes d'indexation et de recherche des vidéos basées sur les ontologies. Une étude comparative de notre système d'indexation OVIS avec ceux de l'état de l'art sera illustrée dans le Tableau 1 ci-dessous. Par la suite, nous allons détailler nos contributions scientifiques. Enfin, nous terminerons le chapitre par une conclusion.

2.2. Indexation et recherche des vidéos

L'indexation a pour but de substituer à une vidéo un représentant (ou un descripteur) moins encombrant qui la caractérise le plus sémantiquement afin de l'utiliser lors de la recherche. Cette méthode permettra une meilleure organisation des données qui limitera la quantité de données examinées durant une recherche, afin d'y accéder rapidement et de confiner la recherche au maximum.

Il existe actuellement plusieurs méthodes d'indexation établies sur des caractéristiques proches de la représentation du signal, à savoir (la couleur, la texture, la disposition de points caractéristiques...). Par contre, il n'existe pas d'algorithmes en général pour interpréter le contenu d'une image ou d'une vidéo. Dans la plupart des cas, il est possible de définir des critères objectifs permettant de classifier automatiquement le contenu d'une vidéo, lui donnant ainsi un sens. Pour la surveillance d'autoroutes par exemple, le problème est restreint par le fait que l'on ne traite que des véhicules sur des images prises par une caméra fixe. Il est alors possible de caractériser de façon non ambiguë ces véhicules, qui apparaissent comme des régions de petite taille sur un fond stable.

Dans le cadre général d'indexation des séquences vidéo, les critères universels d'interprétation sont difficiles à définir vu la diversité des contenus des vidéos. Par exemple, si on considère le cas des séquences vidéo d'un journal télévisé, on voit que les plans de la vidéo se succèdent d'une façon quasiment statique avec des reportages sur le terrain où les mouvements peuvent être rapides, et les objets d'intérêts étant de nature totalement variable. Dans ce cas, un système d'analyse automatique de vidéo ne peut pas fournir les informations sur le contenu sémantique de la vidéo, et par conséquent un problème d'interprétation de vidéo est alors posé. Il faut alors impliquer un spécialiste du domaine (c'est-à-dire un opérateur humain) pour une classification générale des vidéos.

De ce fait, l'indexation manuelle produit des tables d'index représentant explicitement le contenu sémantique des documents (souvent à l'aide de mots-clefs), ce qui représente

généralement un avantage majeur. Cependant la qualité des index manuels dépend de la capacité de l'opérateur à décrire l'ensemble d'une scène par quelques descripteurs dans le temps imparti : ainsi, si chaque scène doit être décrite en peu de temps, la description sera plus superficielle, et donc moins d'informations seront disponible pour rechercher les documents répondant à une requête donnée. Pour mieux comprendre la terminologie liée à la vidéo, nous proposons une comparaison de ses caractéristiques par rapport à d'autres types de documents. Dans [21], différents critères ont été retenus pour spécifier un document vidéo.

- **Volume :** les données vidéo sont trop étendues et nécessitent plus d'espace de stockage par rapport à l'image et au texte. En effet, une seconde de vidéo MPEG contient 25 ou 30 images.
- **Hétérogénéité du contenu :** les informations contenues dans un document vidéo proviennent de sources variées (audio, texte, image). Cette hétérogénéité cause souvent des problèmes au niveau de la segmentation et l'analyse du contenu vidéo.
- **La nature spatio-temporelle :** les données textuelles et images sont statiques et elles sont caractérisées par un aspect spatial. Par contre, une vidéo est caractérisée par un aspect spatio-temporel.
- **L'expressivité sémantique :** il est facile d'élucider le contenu d'un document vidéo en visualisant son contenu et en exploitant l'information sonore associée. Un document vidéo possède une expressivité sémantique beaucoup plus riche que les documents texte et image.
- **Durée :** pour présenter une image fixe, la durée de présentation est variable. Dans un document vidéo, une image bénéficie d'une durée de présentation bien définie qui dépend du type de vidéo (~1/30 seconde pour le cas de vidéo NTSC et 1/25 seconde pour le cas des vidéos PAL / SECAM).
- **Variabilité de la qualité :** la qualité d'une vidéo dépend de la résolution dans laquelle elle est codée ainsi que du taux de compression utilisé pour son codage. Un même contenu peut être codé avec des résolutions très variables en fonction des capacités de stockage et/ou de transmission. Cela a un impact important sur la qualité des traitements qui peuvent être effectués pour une indexation automatique. De ce fait, le cas de la vidéo est différent de celui du texte ou de l'image pour lequel le contenu et son interprétation ne dépendent pas du format dans lequel ils sont encodés.

2.3. Méthodes d'indexation et de recherche vidéo : Etat de l'art

Dans les applications de vidéosurveillance, l'ontologie peut être utilisée dans le processus d'indexation pour la détection des évènements comme une situation d'un comportement anormal ou alors aussi des situations d'embouteillage de voitures depuis des séquences vidéo générées essentiellement par des caméras de vidéosurveillance stationnaires, etc. En effet, une ontologie qui peut à la fois décrire un domaine d'une façons précise et générer le résultat du raisonnement obtenu à partir des règles SWRL, augmente d'une manière impressionnante la précision sémantique du processus d'indexation.

L'approche ontologique a été utilisée dans plusieurs travaux dans la littérature. Kless et al. [22] qui présentent le thesaurus ou les taxonomies comme méthodes les plus performantes pour la création et l'implémentation des ontologies. Atta et al. [23] développent un système basé sur un réseau d'ontologies évolutives qui ont la possibilité d'indexer une grande base de données des effets spéciaux relatifs à la séquence vidéo. Le système proposé permet une recherche intelligente dans le domaine de la post-production des films. Mariano et al. [24] présentent un système qui a la possibilité de répondre aux questions des ontologies incluant plusieurs optimisations spécifiques. D'une part, ils exploitent les avantages des ABox (assertion component) qui représentent généralement les composants d'assertion ou alors les instances de classes. D'autres parts, ils respectent les propriétés des TBox (terminological component) qui décrivent généralement les différents systèmes en termes de vocabulaire contrôlé. Scherp et al. [25] proposent la notion de noyaux d'ontologie comme un système basé sur des notions logiques de réductibilité, plutôt qu'une distinction entre des ontologies du domaine et des ontologies génériques. Benmokhtar et al. [26] utilisent une approche qui intègre le paradigme d'ontologie avec un réseau de neurones pour la détection des concepts. Rector et al. [27] considèrent la création d'annotations utilisant les ontologies existantes comme une bonne pratique. Smith et al. [28] présentent une théorie d'ontologie spécifique appelée « Ontological Realism » pour une création d'ontologie de haute qualité, utilisant à la fois la vision philosophique (l'étude des entités existantes et la façon avec laquelle elles sont reliées entre elles) et la vision informatique (i.e. la conceptualisation d'un domaine). Soner et al. [29] utilisent l'ontologie pour l'extraction des instances depuis un corpus de document, et rajoutent ces derniers à la base de connaissance. Till et al. [30] résolvent le problème d'utilisation d'une variété de langages et proposent un langage d'ontologie distribué DOL qui permet aux

utilisateurs d'utiliser leurs propres formalismes d'ontologie, tout en tenant compte de l'interopérabilité avec les autres ontologies. Ballan et al. [31] reconnaissent les évènements du domaine de vidéo surveillance et des journaux télévisés en intégrant la connaissance au sein de l'ontologie. Bagdanov et al. [32] utilisent une ontologie multimédia qui contient des prototypes visuels représentant chaque cluster avec la possibilité d'agir comme un pont entre le domaine et l'ontologie de la structure vidéo. Ses auteurs présentent un system qui permet la solution au problème de gap sémantique entre le haut niveau conceptuel et le bas niveau des descripteurs. Bertini et al. [33] classifient les évènements et les objets observés dans la vidéo séquence en rajoutant de nouvelles instances de concepts visuels pour leurs ontologies à travers un mécanisme de mise à jour de concepts existants. L'approche utilise à la fois les descripteurs du domaine générique et du domaine spécifique pour l'identification des prototypes visuels représentant les éléments des concepts visuels. Afin de résoudre le problème de la création manuelle des règles d'inférences SWRL (Semantic Web Rule Language) par les experts du domaine, Bertini et al. [33] proposent une adaptation de la technique FOIL (First Order Inductive Learner), appelé FOILS. Xue et al. [34] proposent un système d'archivage et recherche de vidéo surveillance basée sur les ontologies. Lee et al. [35] classifient et indexent les séquences de vidéo surveillance à travers la création d'un système appelé VOS (Video Ontology System). Snidaro et al. [36] utilisent un ensemble de règles SWRL pour la détection des évènements dans le domaine de la vidéo surveillance.

Le problème posé lors de l'utilisation d'une ontologie dans le processus d'indexation vidéo se caractérise par la façon avec laquelle cette ontologie a été créée. Un autre point désavantageux réside dans le fait que plusieurs travaux antécédents ont utilisé l'ontologie comme outils et ont démontré son efficacité à aider et gérer les processus d'indexation et de recherche vidéo. Cependant, ils se sont basés sur des expérimentations d'évènements, considérant seulement un ou deux objets d'intérêt dans la séquence vidéo. Dans le cas d'objet immobile, ils considèrent des évènements tels que des objets abandonnés ou perdus alors que dans le cas d'objets mobiles, ils considèrent des évènements comme une personne qui traverse de gauche à droite, le décollage d'un avion, etc. Le question qui se pose, représente la façon avec laquelle on peut confirmer l'efficacité de l'ontologie dans les processus d'indexation et de recherche vidéo lorsqu'il s'agit de plusieurs objets d'intérêt. On peut citer comme exemple, un groupe de personnes qui marchent, qui courent, etc…

Calavia et al. [37] développent un système intelligent basé sur une Ontology de vidéo surveillance qui a la possibilité d'analyser les mouvements d'objets et qui identifie les situations anormales et alarmantes. Cependant, le domaine d'application couvert par la documentation n'est pas consistant avec la représentation de l'ontologie. Georgios et al. [38] proposent un algorithme générique pour l'optimisation de la taille de chaque élément de l'ontologie (e.g. concepts, etc). Dans leurs approches, ils considèrent l'importance variable de l'information pertinente globale et locale pour la détection des différents éléments de l'Ontologie. Malheureusement, le type de relation some/some est utilisé à la place de all/some. Sawsan et al. [39] construisent une ontologie pour les mouvements vidéo pour l'annotation automatique des mouvements humains en utilisant l'annotation classique de « Benesh ». L'utilisation de la relation « Is-A » présente des anomalies dans un sens non transitif. On peut citer comme exemple la relation entre les deux concepts (media et vidéo) dans la partie représentation multimédia. Ilias et al. [40] proposent une approche d'ontologie bien structurée pour la représentation des évènements décrite sous forme de graphe de connections entre les différents concepts, avec la représentation d'un certain domaine. Malheureusement, leur approche présente quelques limites concernant l'ontologie et son application dans l'analyse vidéo. Bohlken et al. [41] considèrent le problème de l'interprétation haut niveau de la scène en suggérant pour cette fin une architecture nouvelle, basée sur la génération de règle depuis une ontologieOWL-DL. Cependant, le concept « vehicle entering a zone » n'est pas conceptuel, vu qu'il représente une action entre deux concepts « vehicle and zone »et non un seul concept.

Nevatia et al. [42] développent deux langages appelés « VERL » (Video Event Représentation Langage) et « VEML » (Video Event Markup Language), pour la description des évènements d'ontologie « VERL », et l'annotation des instances des évènements « VEML » respectivement. Cependant, une confusion existe entre le langage d'objet qui décrit le domaine et le meta-langage qui définit ce langage d'objet. Bai et al. [43] présentent un system d'analyse de contenu vidéo sémantique basé sur une ontologie. Les concepts de haut niveau sont décrits en référence à ce domaine d'application et combiné au standard MPEG-7 pour l'expression du contenu bas niveau de l'algorithme d'analyse. Malheureusement, cette ontologie surcharge la relation « Is-A », de plus, une confusion existe entre cette relation « Is-A » et la relation « Instance-Of ». Par exemple, la combinaison de plusieurs instances d'algorithmes avec la relation « Is-A » remplace la relation « Instance-Of ». San Miguel et al.

[44] proposent une approche basée sur une ontologie pour représenter la connaissance de base de l'analyse des évènements vidéo, constituée de deux types différents de connaissances : le domaine d'application et le système d'analyse. Le domaine d'application inclue tous les concepts de haut niveau (objets, évènements, contexte, etc.), pendant que le système d'analyse inclue les compétences de ce dernier (algorithmes, réactions pour les évènements, etc.). Cependant, cette ontologie détermine seulement le meilleur module d'analyse, et elle n'inclue aucune inférence pour la détection des évènements, en plus du fait que cette ontologie surcharge la relation « Is-A ».

En respectant le fait qu'on peut distinguer trois étapes dans le processus d'indexation : bas niveau, niveau intermédiaire, et haut ou niveau sémantique, différentes approches peuvent être trouvées dans la littérature utilisant les descripteurs dans le bas et niveau intermédiaires et les classifieurs dans le niveau sémantique ou haut niveau. Utasi et al. [45] proposent une approche basée sur des descripteurs statistiques pour la détection de trois types d'évènements : activité régulière, courir, et séparation de groupes. Cette approche commence par l'utilisation d'une technique d'extraction d'arrière-plan suivie par le calcul de flux optique pour les pixels d'avant plan. Cependant, cette approche ne détecte pas un grand nombre d'évènement comme la marche ou la formation de groupe. Chan et al. [46] utilisent un model basé sur les propriétés globales pour la détection des évènements comme ceux de la marche, course, formation, ou séparation. Leur approche caractérise le flux de la foule en utilisant une texture dynamique. Malheureusement, ce modèle n'inclue pas la possibilité de gérer deux évènements à la fois comme par exemple l'évènement de marche (commence à l'image 000, et se termine à l'image 310) et l'évènement formation (commence à l'image 191, et se termine à l'image 340).

D'autres approches sémantiques ont participé au challenge TRECVID 2016 [47]. Markatoupoulou et al. [48] proposent un system de détection d'évènement de vidéo-surveillance basé sur la méthode d'encodage Fisher Vector et en suivant le modèle SVM pour étudier comment séparer chacune des activités. Cependant, le point négatif de cette approche réside dans le fait que celle-ci retourne plusieurs fausses alarmes. Zhao et al. [49] utilisent plusieurs approches à la fois pour la détection d'évènements et leurs systèmes se divise en deux grandes parties : rétrospective et interactive. La première partie détecte les piétons, effectue leurs suivis et détecte l'évènement adéquat. La seconde partie corrige la mauvaise détection d'évènement. Malheureusement, cette méthode ne détecte que peu d'évènements.

2.4. Etude comparative des travaux existants dans la littérature

Notre contribution [50] dans le Workshop intitulé : "**Computer Vision + ONTology Applied Cross-Disciplinary Technologies**", qui se tient tous les deux ans en alternance avec la conférence internationale sur la vision par ordinateur ECCV (European Conference on Computer Vision), qui s'est déroulé à Zurich en Suisse les 6-12 Septembre 2014, un des Workshops le plus référencé dans le domaine d'indexation et de recherche vidéo, nous a permis d'établir une étude comparative de notre ontologie de vidéosurveillance avec celles existantes dans la littérature (voir tableau 1).

Métriques / Ontologies	Consistante	Formelle	Largeur du Domaine de représentation	Conceptualisation	Séparation des deux relations IS-A et INSTANCE-OF	Utilisation d'inférence (SWRL)
OVIS	X	X	X	X	X	X
Calavia et al. [37]		X	X		X	X
Sawsan et al. [39]	X		X	X		X
Trochidis et al. [40]	X	X		X	X	X
Bohlken et al. [41]	X		X		X	X
Bai et al. [43]		X	X	X		X
SanMiguel et al. [44]	X	X		X	X	

Tableau 1. Etude comparative de notre ontologie de vidéosurveillance avec celles existantes.

Notre ontologie proposée dans cette thèse utilise une syntaxe formelle et sémantique pour la consistance et vérifie toutes les métriques standards. Elle est complète, du fait qu'elle modélise grâce aux règles d'inférences (SWRL) la majorité des événements existants dans le domaine académique ou industriel de la surveillance.

2.5. Les contributions essentielles de la thèse

Dans cette thèse, nous avons conçu et réalisé un système d'indexation et de recherche vidéo appelé OVIS qui utilise une ontologie de vidéosurveillance et des règles SWRL [6, 50, 51]. Ainsi, nos différentes contributions dans cette thèse peuvent être résumées comme suit :

- Un état de l'art détaillé des approches et des systèmes d'indexation et de recherche des vidéos nous a permis :
  - d'identifier les inconvénients et les limites des approches existantes de la littérature.

- d'établir un tableau comparatif de notre ontologie de vidéosurveillance avec celles existantes dans la littérature selon les métriques suivantes : la consistance, le formalisme, la largeur du domaine de représentation et la conceptualisation.

- La création de l'ontologie de vidéosurveillance suit une convention de nommage syntaxique bien définie afin d'assurer sa consistante et sa complétude.
- L'ontologie de vidéosurveillance proposée est une extension de celle de San Miguel et al. [44] et modélise tous les concepts du domaine de la vidéosurveillance. En effet, nous avons ajouté de nouveaux concepts utilisés dans la surveillance industrielle comme par exemple l'évènement d'intrusion.
- Notre ontologie de vidéosurveillance se distingue des autres de la littérature par la diversité des cas de son utilisation, comme par exemple :
  - La création des benchmarks,
  - La représentions des scènes où
  - La représentation des vidéos en général.
- Notre ontologie de vidéosurveillance utilise des règles d'inférences SWRL aux niveaux moyens et haut de la hiérarchie d'indexation des vidéos.
- Notre système d'indexation et de recherche des vidéos appelé OVIS et que nous avons conçu et réalisé durant les années de thèse. En outre, l'expérimentation et l'évaluation d'OVIS en utilisant deux benchmarks universels tels que PETS 2012 et TRECVID 2016

2.6. Conclusion

De nos jours, il existe assez peu de travaux de recherche qui traitent les besoins des utilisateurs car ils sont basés sur les aspects physiques comme l'analyse et l'extraction automatique de l'information audiovisuelle. Cependant, Dans un contexte de recherche d'information vidéo sémantique, la modélisation représente une tâche très importante et surtout nécessaire au processus de l'indexation, afin de rendre la recherche vidéo plus efficace et précise. De cette manière, l'intégration d'une ontologie dans les systèmes d'indexation et de recherche vidéo, se présente comme une tâche de modélisation nécessaire aux besoins des utilisateurs.

En informatique, une ontologie représente un ensemble structuré de concepts permettant de donner un sens aux informations. Son rôle principal est de modéliser un ensemble de connaissances dans un domaine donné. Afin d'ajouter une couche sémantique aux systèmes informatiques, les ontologies se présentent comme un outil de représentation des descriptions

formelles du domaine. Par conséquent, les ontologies sont employées dans plusieurs thématiques de recherche en informatique tels que l'intelligence artificielle, le web sémantique, le génie logiciel, l'informatique biomédicale etc. ….

Dans ce chapitre, nous avons présenté le domaine d'indexation et de recherche vidéo ainsi que ses caractéristiques. Ensuite, nous avons présenté les avantages et les inconvénients des méthodes d'indexation utilisant les ontologies. Par la suite, nous avons dressé un tableau comparatif de notre ontologie de vidéosurveillance avec celles de la littérature et nous avons expliqué nos différentes contributions de thèse.

Dans le chapitre 3 suivant, nous allons détailler la solution proposée d'indexation et de recherche des vidéos dans une vidéosurveillance.

Chapitre III : Solution Proposée basée sur l'approche ontologique

3.1. Introduction

Un très grand nombre de chercheurs ont porté un grand intérêt quant à l'indexation des vidéos d'une vidéosurveillance pour l'analyse du comportement humain et la majorité des travaux de recherche concerne l'interprétation sémantique du mouvement humain. Pour cela, des méthodes d'extraction d'information de bas niveau sont utilisées comme par exemple les blobs afin de résoudre le problème de la sémantique relative à une séquence vidéo et d'en déduire alors le comportement correspondant. Pour aboutir à cette fin, l'approche ontologique est perçue comme un outil efficace de modélisation, d'indexation et de recherche des vidéos. Dans ce qui suit, nous allons tout d'abord introduire les bases théoriques des ontologies. Ensuite, nous allons détailler notre ontologie de vidéosurveillance à savoir sa syntaxe et sa sémantique ainsi que les domaines de ses applications tels que : la création des benchmarks, la description des scènes, ….

3.2. Domaine des Ontologies

Une ontologie en informatique est un ensemble structuré de concepts permettant de donner un sens aux informations. Son premier objectif est de modéliser un ensemble de connaissances dans un domaine donné. De plus, les ontologies informatiques sont des outils qui permettent précisément de représenter un corpus de connaissances sous une forme utilisable par un ordinateur.

Un des enjeux courants de l'indexation et de la recherche vidéo est celui d'accroître des systèmes capables d'inclure plus de sémantique dans leurs traitements. L'objectif est double : « comprendre » les contenus des documents vidéo et « comprendre » le besoin de l'utilisateur pour pouvoir les mettre en relation.

Les ontologies sont utilisées pour représenter des descriptions partagées plus ou moins formelles du domaine et d'ajouter ainsi une couche sémantique aux systèmes informatiques. Il est donc naturel que des travaux sur l'intégration des ontologies dans les systèmes d'indexation et de recherche de vidéo se développent. Nous situons cette partie dans ce cadre-là. Les ontologies sont employées dans l'intelligence artificielle, le web sémantique, le génie logiciel, l'informatique biomédicale et l'architecture de l'information comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde.

Nous allons commencer par les bases théoriques des ontologies. Nous évoquerons ensuite les différents composants constituant une ontologie, les différents types d'ontologies avant de donner les langages de représentation d'ontologies.

3.2.1. Bases théoriques

a. Définition d'une ontologie

La définition la plus référencée et aussi la plus synthétique est sans doute celle de Gruber [52] : une ontologie est une spécification explicite d'une conceptualisation. Cette définition est étoffée dans [53]. Une ontologie est définie comme étant un ensemble de définitions et de représentation de connaissances spécifiques au contenu : classes, relations, fonctions et constantes d'objet. La même notion est également développée dans [54] : une ontologie est une théorie logique avec une base de connaissance particulière, dont les modèles imposent une certaine conceptualisation.

Les concepts sont agencés dans un graphe dont les relations peuvent être :

- Des relations sémantiques.
- Des relations de composition et d'héritage (au sens objet).

Le terme d'ontologie est utilisé depuis le début des années 1990 dans le domaine de l'intelligence artificielle (IA), en particulier de l'ingénierie des connaissances et de la représentation des connaissances. Son champ d'application s'agrandit abondamment et il fait dorénavant partie des objets de recherche courants. Une ontologie est un système formel dont l'objectif est de représenter les connaissances d'un domaine spécifique au moyen d'éléments de base, les concepts définis et organisés les uns par rapport aux autres.

Le Web sémantique a besoin d'ontologies ayant un degré de structure révélatrice. C'est pourquoi on doit définir des descriptions pour les concepts suivants :

- **Individus** : les objets de base.
- **Classes :** ensembles, collections, ou type d'objets.
- **Attributs :** propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager.
- **Relations :** les liens que les objets peuvent avoir entre eux.
- **Evénements :** changement subis par des relations ou des attributs.

Une ontologie est un exemple d'organisation des connaissances dans un domaine donné. Par exemple, dans une ontologie du domaine de la vidéosurveillance, on trouve les classes d'objets à organiser (Person, Group_Of_Person, Bag, Video_Sequences ...), les types d'attributs

qui peuvent être attachés aux objets (Velocity, Person_Number, StartPosition, Bag_Number...) et les types de relations entre les objets (un objet "Person" peut être relié par une relation "Occure_In" à un objet de type "Video_Sequences"), etc.

b. Les objectifs de l'ontologie

L'ontologie peut servir pour divers objectifs parmi lesquels on peut citer :

- La communication (humains et organisations) : Dans l'ontologie, il n'y a jamais deux termes ayant la même sémantique. Cette situation se produit souvent si l'on utilise un langage naturel pour la communication.
- L'interopérabilité (machine et systèmes) : l'ontologie sert à définir le format d'échange entre les systèmes.
- L'ingénierie des systèmes : l'ontologie peut servir divers aspects du développement des systèmes d'information. Elle assiste au processus de construction de la spécification des systèmes. Elle soutient aussi l'automatisation du processus de vérification de la fiabilité des systèmes.

3.2.2. Le processus de création d'une ontologie

Le processus de création d'une ontologie doit être considéré comme un projet. Les méthodes de gestion de projet peuvent donc s'appliquer à ce processus**.**

Le cycle de vie des ontologies est inspiré du génie logiciel. Il comprend une étape initiale d'évaluation des besoins, une étape de construction, une étape de diffusion, et une étape d'utilisation. Après chaque utilisation significative, l'ontologie et les besoins sont réévalués et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite. Ainsi, la construction se compose de quatre étapes :

a. Spécification

La construction d'une ontologie commence par la définition d'un domaine et de sa portée. C'est à dire, il faut trouver des réponses à des questions comme [55] :

- Quel est le domaine que l'ontologie abreuvera ?
- Quels sont les buts de l'utilisation de l'ontologie ?
- A quels types de questions l'information contenue dans l'ontologie devra-t-elle fournir des réponses ?
- Qui va utiliser et maintenir l'ontologie ?

b. Conceptualisation

Identification des connaissances contenues dans un corpus représentatif du domaine. Ce travail doit être mené par un expert du domaine, assisté par un ingénieur de la connaissance.

c. Formalisation

Une fois le modèle conceptuel organisé, il faut le traduire dans un formalisme. Grâce à la formalisation, les définitions des concepts sont plus explicites et précises. L'objectif est de faciliter l'interprétation de l'ontologie. Uschold et Jasper [56] ont soumis quatre niveaux de formalismes d'une ontologie, du haut niveau informel au niveau formel ascétique. Le niveau du formalisme d'une ontologie est choisi selon les besoins et selon le langage d'implémentation de l'ontologie. Par exemple, si l'ontologie est un Framework pour la communication entre des personnes, alors la représentation d'une ontologie peut être informelle. Cependant, si l'ontologie est utilisée par les outils logiciels, la représentation doit être plus formelle.

Parmi les nombreux langages de formalisation des ontologies, il y a trois grandes familles : les langages à base de frames, les modèles de graphes contextuels et les logiques de description.

Dans les langages à base de frames, les frames représentent les catégories d'objets et sont dotées d'attributs (slots). Les slots peuvent prendre différentes valeurs. Une instance de classe, qui a un identifiant et des attributs uniques est liée à sa classe par le lien « is-a ». Les classes sont structurées par un lien hiérarchique « a-kind-of ». Flogic [57] est un exemple connu de formalisme qui intègre les langages à base de frames, les formalismes orientés objet et la logique du premier ordre.

Dans le modèle des graphes conceptuels [58], on caractérise des niveaux différents. Le niveau conceptuel, peut servir de base à un langage spécialisé de communication entre les spécialistes de différentes disciplines impliquées dans un travail cognitif commun. Le niveau d'exécution peut servir de base à un outil commun de représentation employé par plusieurs modules d'un système complexe.

Les logiques de description sont basées sur la logique des prédicats, les réseaux sémantiques et les langages à base de frame. Dans le formalisme des logiques de description, les connaissances sont exhibées sous forme de concepts, d'individus et de rôles. Un concept est une entité générale d'un domaine d'application. Les rôles sont des relations binaires entre concepts et les individus sont les instances de concepts. Les propriétés des concepts, rôles et individus sont exprimées en logique des prédicats.

d. Implémentation

En résumé, il faut qu'on implémente l'ontologie dans un langage. Le langage choisi doit correspondre au modèle de formalisation. Une introduction brève des langages d'ontologie est présentée ultérieurement.

3.2.3. Composants des ontologies

Les connaissances traduites par une ontologie sont véhiculées à l'aide de cinq éléments [59] : Concepts ; Instances ; Fonctions ; Axiomes ; Relations.

a. Les concepts

Ils sont appelés aussi termes ou classes de l'ontologie. Un concept est un élément de la pensée (un principe, une idée, une notion abstraite) sémantiquement évaluable et communicable. Ces concepts peuvent être classifiés selon plusieurs dimensions :

- Niveau d'abstraction (concret ou abstrait) ;
- Atomicité (élémentaire ou composée) ;
- Niveau de réalité (réel ou fictif). Pour résumé, un concept peut être tout ce qui peut être évoqué : description d'une tâche, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement, etc.

b. Les relations

Les associations existantes entre les concepts présents dans le segment analysé de la réalité sont traduites par elles. Ces relations regroupent les associations suivantes : sous-classe-de (spécialisation, généralisation) ; partie de (agrégation ou composition) ; associée-à ; instance-de ; est-un, etc. ces relations nous permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres. Les relations représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est à dire R : C1*C2*……Cn.

c. Les fonctions

Ce sont des cas distinctifs de relations dans lesquelles le énième élément de la relation est défini de manière unique à partir des n-1 premiers.
Formellement, les fonctions sont définies ainsi : F : C_1*C_2*…*Cn-1→ C_n.

d. Les axiomes

Les axiomes sont utilisés pour ordonnancer des phrases qui sont toujours vraies. Ils constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine traduites par l'ontologie.

e. Les instances

Elles sont utilisées pour la représentation de certains éléments.

3.2.4. Les différents types d'ontologies

Les ontologies peuvent être classifiées selon plusieurs dimensions. Nous en examinerons les deux suivantes :

a. Objet de conceptualisation.

Les ontologies sont classifiées selon leur objet de conceptualisation (c'est-à-dire selon le but de leur utilisation) comme ci-dessous :

- Haut niveau/générique ;
- Domaine ;
- Tâche ;
- Application.

La Figure 7 en dessous montre les différents types d'ontologies selon leur objet de conceptualisation.

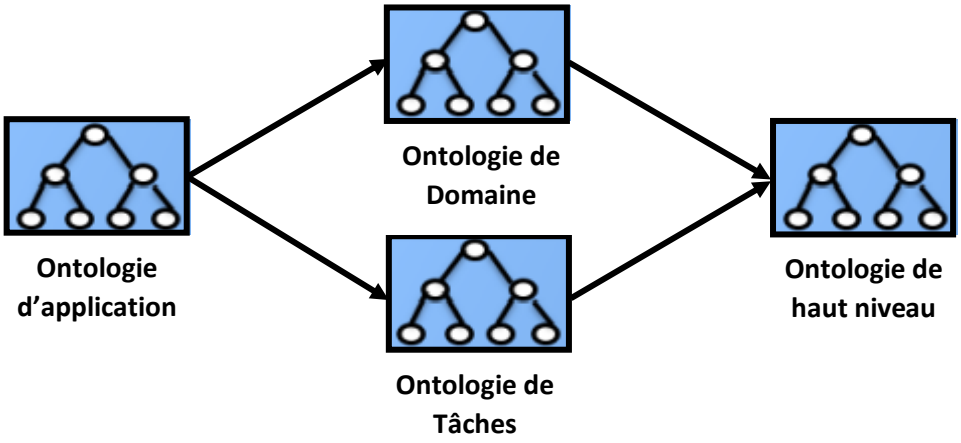


Figure 7. Les types d'ontologies selon leur objet de conceptualisation.

➢ **Les ontologies de haut niveau ou génériques**

Elles expriment des conceptualisations valables dans déférents domaines : décrivent des concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc. Ces concepts ne dépendent pas d'un problème ou d'un domaine particulier, et

doivent être, du moins en théorie, consensuels à de grandes communautés d'utilisateurs. Des exemples d'ontologies de haut niveau sont Dolce ou Sumo.

- **Les ontologies de domaine**

Elles sont réutilisables par plusieurs applications sur ce domaine à l'opposé des ontologies de haut niveau. Elles synthétisent les connaissances distinctes à un domaine particulier. Elles décrivent le vocabulaire ayant trait à un domaine générique (ex. : l'enseignement, la médecine…), notamment en spécialisant les concepts d'une ontologie de haut niveau.

- **Les ontologies de tâches**

Ce type d'ontologies est utilisé pour conceptualiser des tâches distinctes dans les systèmes, comme par exemple les tâches de diagnostic, de planification, de conception, de configuration, de tutorat. Soit tout ce qui concerne la résolution de problèmes. Ce type d'ontologies décrit le vocabulaire concernant une tâche générique (ex. : enseigner, diagnostiquer…), particulièrement en spécialisant les concepts d'une ontologie de haut niveau. Quelques auteurs adoptent le nom « ontologie du domaine de la tâche » pour faire référence à ce type d'ontologie.

- **Les ontologies d'application**

Elles contiennent des connaissances du domaine nécessaire pour une application donnée ; spécifique, non réutilisable. C'est l'une des ontologies les plus spécifiques. Ces concepts concordent souvent aux rôles joués par les entités du domaine lors de l'exécution d'une certaine activité. Ici, il s'agit donc de mettre en relation les concepts d'un domaine et les concepts liés à une tâche particulière, de manière à en décrire l'exécution (ex. : apprendre les statistiques, effectuer des recherches dans le domaine de l'astronomie, etc.).

b. Niveau de formalisme de représentation

Les langages dédiés aux ontologies sont notamment résultant des formalismes liés aux réseaux sémantiques, les graphes conceptuels, les frames et les logiques de description.

- **Les réseaux sémantiques**

Un réseau sémantique est une représentation graphique d'une conceptualisation d'une (ou plusieurs) connaissances humaines [60]. Il est figuré sous la forme d'un graphe orienté et étiqueté ou plus expressément un multi graphe car deux nœuds du graphe peuvent être reliés par plusieurs arcs.

Un ensemble de nœuds typés, dénotant des concepts du domaine modélisé, et d'arcs orientés étiquetés le constitue, représentant les relations sémantiques entre les concepts. Ainsi un concept est décrit par les autres concepts du réseau en relation avec le premier. Certains nœuds correspondent intuitivement mieux à des classes d'objets qu'à des individus. Représenter l'appartenance à une classe nécessite une relation d'appartenance ; c'est pourquoi les réseaux possèdent un nom réservé d'étiquette pour cette relation, parfois nommé ''sorte de''. Mais cette relation d'appartenance à une classe suppose que l'on sait différencier les nœuds qui représentent des classes de ceux qui dénotent des individus.

De nombreux formalismes de représentation des connaissances dérivés des réseaux sémantiques imposent de noter différemment les deux types de nœuds (par exemple KL-ONE [61]). La relation d'appartenance à une classe permet de rapporter les connaissances de la classe sur un individu.

- **Les graphes conceptuels**

Un graphe conceptuel est un formalisme de représentation de connaissances et de raisonnements. Ce formalisme à base de graphe a été introduit par John F. Sowa en 1984 [62]. Un graphe conceptuel est un multi graphe, il contient deux sortes de nœuds : les nœuds concepts, qu'on appelle aussi sommets concepts ou plus sommairement concepts, et les nœuds relations ou relations. Chacun de ces nœuds à une étiquette. Un nœud concept est étiqueté par un type correspondant à une classe sémantique, et un marqueur précisant une instance particulière de classe. Les nœuds relations sont aussi étiquetés par un type (voir Figure 8).

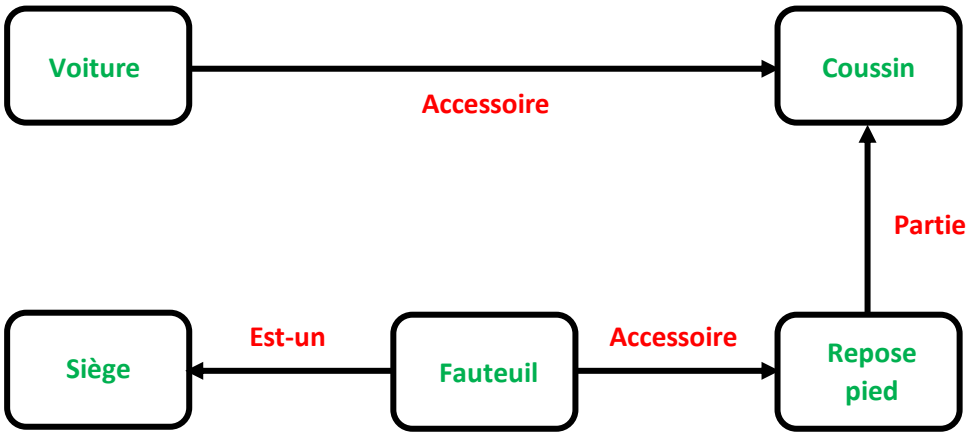


Figure 8. Exemple d'un graphe conceptuel.

- **Les frames**

Dans les langages de frames, les connaissances sont ameutées en paquets. C'est pourquoi les relations d'un concept du domaine avec les autres concepts font partie de la description de

ce concept. Dans les réseaux sémantiques, il se trouve une unité sémantique dénotant un concept (le représentant), et un graphe composé d'autres unités sémantiques qui décrit la définition de ce concept. Dans un frame l'unité sémantique dénotant le concept, est fusionné avec sa description, c'est-à-dire qu'une seule unité sémantique comprend toute la description du concept et est aussi utilisée pour représenter ce concept.

En 1975, Minsky propose le premier formalisme informatique à partir des idées précédentes. Le frame correspond à une structure dynamique représentant des situations prototypages. Schank, qui s'intéresse à la représentation de séquence d'évènements et la compréhension d'histoires, aboutit quasiment au même résultat avec la théorie des scripts [63].

Un frame est un exemplaire, c'est à dire un objet particulier d'une famille, représentant parfaitement cette famille. Le frame contient donc des informations générales valides pour tous les membres de la famille, ainsi que des informations caractéristiques à certains membres.

Les frames sont représentés par une structure de données capable de mettre en valeur des objets structurés.

3.2.5. Les langages d'ontologie

Pour implémenter des ontologies, plusieurs langages ont été utilisés. La plupart d'entre eux sont basés sur XML et les logiques de description. Dans cette section, nous présenterons brièvement XML, XML Schema, RDF, RDF Schema et OWL qui sont les langages les plus connus.

a. Le langage XML

XML (eXtensible Markup Language) recommandé par le W3C (World Wide Web Consortium) est une spécification destinée à rendre des documents lisibles par une machine.

XML fournit seulement une structure syntaxique pour des documents et ne permet pas une interprétation sémantique des données [64].

XML Schema qui est recommandé par le W3C, permet de définir les balises ainsi que l'agencement de ces balises autorisé pour définir la validité d'un document XML.

b. Le langage RDF

Il a été créé par Tim Berners Lee, au début des années 1990, le web était essentiellement destiné à partager des informations sous l'aspect de pages html, affichables par un logiciel « navigateur web », et fréquemment destinées à être lues par un utilisateur humain.

Très rapidement, on s'est rendu compte que cette conception du web était bien trop abrégée, et ne permettait pas un réel partage du savoir : tout au plus cela permettait-il de présenter des connaissances, mais en aucun cas de les rendre directement utilisables.

L'arrivée de XML, en 1998, a donné un cadre à la structuration des connaissances, ce qui a rendu ainsi possible la création de nouveaux langages web destinés non seulement à un rendu graphique à l'écran pour un utilisateur humain, mais à un réel partage et à une manipulation des savoirs. C'est dans cet esprit qu'a été créé en 1999 RDF, un langage XML permettant de décrire des métadonnées tout en facilitant leur traitement.

Le développement de RDF par le W3C a été entre autres stimulé par la perspective des applications suivantes :

- Maximalisation de la coopération entre applications, en autorisant de combiner les données de plusieurs applications, pour engendrer de nouvelles informations.
- Développement de modèles d'information ouverts plutôt que fixés pour quelques applications (par exemple les activités de description, de planification de processus organisationnels, d'annotation de ressources web, etc.).
- Établir avec l'information accommodante par machine ce que le Web a fait pour l'hypertexte, en permettant aux informations d'être arrangées en dehors de l'environnement distinctif dans lequel elles ont été créées, hypothétiquement à l'échelle d'Internet. C'est le concept d'interopérabilité des savoirs.
- Manipulation et archivage des métadonnées Web, dans le but de fournir des informations sur les ressources Web et les systèmes qui les utilisent.
- Simplifier le traitement automatique de l'information du Web par des agents logiciels, transformant ainsi le web d'un regroupement d'informations intégralement destinées aux humains, en un état de réseau de processus en coopération. Dans ce réseau, le rôle de RDF est de fournir une lingua franca compréhensible par tous les agents.

Pour ce faire, RDF procède à une description de savoirs (données tout comme métadonnées) à l'aide d'expressions de structure fixée. En effet, la structure fondamentale de toute expression en RDF est une collection de triplets, chacun composé d'un sujet, un prédicat et un objet. Un ensemble de tels triplets est appelé un graphe RDF. Ceci peut être illustré par un diagramme composé de nœuds et d'arcs dirigés (voir Figure 9), dans lequel chaque triplet est représenté par un lien nœud-arc-nœud (d'où le terme de "graphe"). [65]

Triplet RDF

Figure 9. Exemple d'une description RDF.

Dans un graphe, l'existence d'une relation entre les nœuds qui sont joints est représentée par un triplet.

c. Le langage RDFS

Comme on peut le constater dans l'exemple précédent, la propriété « auteur » n'a de sens pour décrire la ressource « http://www.lacot.org/ » que dans un contexte bien défini :

- L'information communiquée par le triplet RDF « {http://www.lacot.org/, Xavier Lacot, auteur} » est convenablement triviale pour comprendre que sa signification est « Xavier Lacot est l'auteur de http://www.lacot.org/ ».
- Ce lecteur comprend le français.
- Le lecteur (utilisateur) est humain.

Il est donc nécessaire, pour donner un sens aux informations stockées sous forme de triplets RDF, de se donner un vocabulaire, de définir la signification de la propriété « auteur », ainsi que son type, son champ de valeurs etc. C'est le rôle de RDF Schema, qui permet de créer des vocabulaires de métadonnées. [66]

d. OWL (Ontology Web Language)

L'étymologie du langage OWL est le World Wide Web Consortium (W3C) qui a mis sur pieds, en Novembre 2001, le groupe de travail « WebOnt », chargé d'étudier la création d'un langage standard de manipulation d'ontologies web. Le premier Working Draft « OWL Web Ontology Language 1.0 Abstract Syntax » paraît en Juillet 2002 et, au final, OWL devient une Recommandation du W3C le 10 Février 2004 ;

OWL est généré pour détailler et représenter un domaine de connaissance particulier, en définissant des classes de ressources ou objets et leurs relations ; ainsi que de définir des individus et affirmer des propriétés les concernant et de raisonner sur ces classes et individus. C'est aussi un standard qui se base sur la logique de descriptions. Il est construit sur RDF et RDFS et utilise la syntaxe RDF/XML.

OWL est prédéterminé à être utilisé quand l'information contenue dans les documents doit être cultivé par des applications, par discordance aux situations où le contenu doit seulement

être présenté. Il est aussi employé pour représenter clairement la signification des termes dans les vocabulaires et les relations entre ces termes.

Cette représentation des termes et leurs affinités s'appellent une ontologie. OWL offre plus de facilités pour exprimer la signification et la sémantique que XML, RDF et RDF-S. OWL va ainsi au-delà de ces langages dans sa capacité de représenter le contenu compréhensible par une machine sur le Web.

OWL est une révision du langage d'ontologie du Web DAML+OIL intégrant les leçons apprises de la conception et de l'application de DAML+OIL.

Les langues ont été utilisées plutôt pour développer des outils et des ontologies mais elles n'ont pas été définies pour être compatibles avec l'architecture du WWW en général et le Web Sémantique spécialement. OWL en fondant sur RDF nous donne les possibilités suivantes aux ontologies :

- Concordance avec des prescriptions du Web pour l'accessibilité et l'internationalisation.
- Capacité d'être distribué à travers beaucoup de systèmes.
- Ouverture et extensibilité.

Jusqu'à présent, il y a pas mal d'organismes qui utilisent OWL avec de nombreux outils disponibles. Aujourd'hui, la plus grande majorité des systèmes qui ont utilisé DAML, OIL, ou DAML+OIL adoptent maintenant OWL. En outre, un certain nombre d'outils de langue d'ontologie, par exemple, Protégé nous donne l'appui pour OWL. De plus, il y a beaucoup d'ontologies disponibles sur le Web qui sont créées par OWL. Par exemple dans la bibliothèque de DAML, on peut utiliser les ontologies pour capturer la connaissance dans le domaine d'intérêt. Une ontologie va décrire les concepts dans ce domaine et les liens entre eux. Les différentes langues d'ontologie ont des avantages différents. En ce moment, OWL est considéré par le W3C comme une langue d'ontologie standard. Il a non seulement la capacité de décrire les concepts dans un domaine mais aussi utilise un ensemble plus riche d'opérateurs. On peut construire des concepts complexes en se basant sur les définitions des concepts plus simples. En outre, on peut vérifier si tous les rapports et les définitions dans l'ontologie sont conformés et par conséquent identifier quels concepts s'adaptent sous quelles définitions. Donc, on peut maintenir la hiérarchie correctement entre les classes. OWL peut faire un compromis entre son pouvoir expressif et son pouvoir de raisonnement parce qu'il fournit des sous langages de plus en plus expressifs conçu. Ainsi OWL permet d'augmenter le sens du vocabulaire prédéfini dans une ontologie. On peut définir alors chaque sous-langage grâce au son expression [67]. Par exemple : OWL-Lite est le moins expressif et OWL-Full est le plus expressif. Par ailleurs, On

peut considérer OWL-DL comme une extension d'OWL-Lite et OWL-Full comme une extension d'OWL-DL.

Les ontologies OWL se présentent, généralement, sous forme de fichiers texte et de documents OWL. Pour cela, OWL offre trois sous langages d'expression (voir Figure 10 ci-dessous), conçus pour des communautés de développeurs et utilisateurs spécifiques :

- Le langage OWL DL implique les utilisateurs qui souhaitent une expressivité maximum sans sacrifier la complétude de calcul (inférences) et la décision des systèmes de raisonnement. Le langage OWL DL contient toutes les structures de langage d'OWL avec des restrictions comme la séparation des types (une classe ne peut pas être en même temps un individu ou une propriété, une propriété doit être un individu ou une classe).
- Le langage OWL Lite concerne les utilisateurs ayant principalement besoin d'une hiérarchie de classifications et de mécanismes de contraintes simples. Par exemple, quoiqu'OWL Lite gère des contraintes de cardinalité, il ne permet que des valeurs de cardinalité de 0 ou 1.
- Le langage OWL Full concerne les utilisateurs souhaitant une expressivité maximum et la liberté syntaxique de RDF sans garantir le calcul (raisonnement). Par exemple, dans OWL Full, on peut simultanément traiter une classe comme une collection d'individus et comme un individu à part entière. Une autre différence significative par rapport à OWL DL réside dans la possibilité de marquer un objet OWL : DatatypeProperty comme étant un objet OWL : InverseFunctionalProperty.

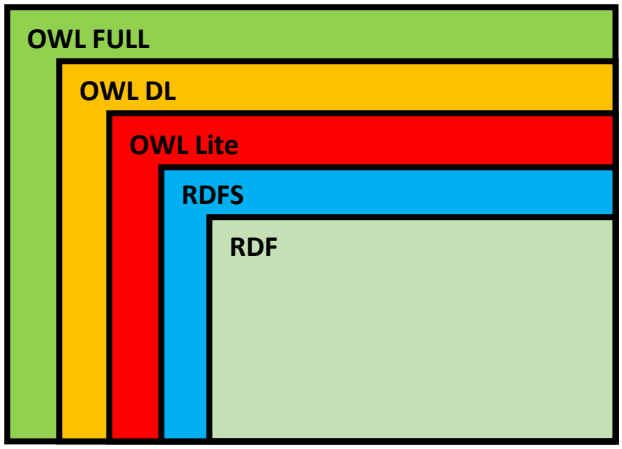


Figure 10. Hiérarchie des sous-langages d'OWL.

3.2.6. Les ontologies d'évènements vidéo

a. Conceptualisation du domaine

Le concept spatial le plus courant est l'*objet physique*. Les objets physiques sont tous des objets du monde réel dans une scène surveillée. Conformément à la capacité de prévision du mouvement des objets, nous définissons deux types d'objets physiques : *mobiles* et *contextuels*.

- Objets mobiles : C'est un objet dont le mouvement ne peut être prévu. Nativement, un objet mobile baptise son mouvement. Des objets mobiles représentatifs sont des individus, des groupes de personnes, des robots, et/ou des animaux.
- Objets contextuels : C'est un objet dont le mouvement peut être anticipé à l'utilisation de l'information à priori. En général, un objet contextuel ne peut pas changer d'endroit dans la scène. Toujours est-il qu'il peut être déplacé par un autre objet. Les objets contextuels typiques sont des murs, des zones d'entrée, des portes, des chaises.

Il y a plusieurs niveaux de granularité pour estimer les objets physiques (Figure 11). A un niveau de granularité très important, un groupe de personnes peut être considéré comme un objet mobile. A un niveau de granularité beaucoup plus fin, une personne peut être considérée comme un objet mobile. A un niveau de granularité encore plus fin, nous pouvons prendre une personne comme une entité complexe qui est capable d'effectuer des actions simultanées avec les parties différentes de son corps. Dans ce cas, chacune des parties du corps peut être vue comme un objet mobile. Ainsi, un objet physique peut induire la création de plusieurs nouveaux objets ou fusionner avec d'autres objets pour former un objet unique.

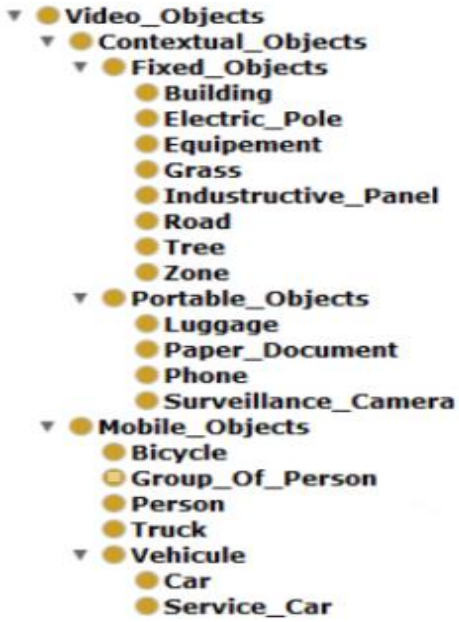


Figure 11. Exemples d'objets physiques.

- **Les relations spatiales**

Les relations spatiales caractérisent des relations entre les objets physiques impliqués dans un événement donné. Il y a deux types importants de relations spatiales : les relations de distance et les relations topologiques.

Les relations topologiques et de distance détaillent des relations entre les objets physiques (par exemple si une personne est proche ou loin d'un distributeur automatique de billets). Les relations de distance peuvent être quantitatives. Par exemple, on peut définir que la distance entre deux personnes est de 5 mètres. On peut aussi définir que l'objet A est proche de l'objet B si la distance entre A et B est moins de 1 m.

- **Les relations temporelles**

Les relations temporelles sont adoptées pour définir les événements. Les relations temporelles englobent des opérateurs d'algèbre d'intervalle d'Allen et des relations quantitatives entre la durée, le commencement et la fin des événements.

L'ontologie de DAML-Time est un langage riche de caractérisation pour l'expression dans une sémantique temporelle, pour les relations topologiques entre des instants et des intervalles ainsi que des relations temporelles.

Les instants sont, intuitivement, un point de temps qui n'a aucun point d'intérieur, et les intervalles sont, intuitivement, des choses avec l'affluence. Ceci signifie que des événements d'instant sont instantanés, comme l'occurrence d'un accident de voiture, d'événements d'intervalle durent sur un certain intervalle, par exemple, une réunion de 2pm à 3pm.

Les relations temporelles auxquelles on s'intéresse pour représenter des événements vidéo sont : avant, après, pendant et rencontre.

- **Etats et événements vidéo**

Les différentes manières pour distinguer les évolutions et les interactions entre objets mobiles dans une scène sont des états et des événements (primitif et composé). Ces types de concepts sont définis ci-dessous.

Un **état** est une propriété spatio-temporelle valide à un instant donné ou stable sur un intervalle de temps. Des objets physiques ont des propriétés ou des attributs en relation avec d'autres objets. Les propriétés, les attributs et les relations peuvent être conçus comme des états, par exemple : une personne est à l'intérieur ou à l'extérieur d'une chambre.

Un **état primitif** est une propriété spatio-temporelle valide et équilibrée sur un intervalle de temps.

Un **état composé** est une combinaison d'états. Nous appelons **composants** tous les sous-états d'un état et nous appelons des **contraintes** toutes les relations concernant ces composants.

Un **événement** est un ou plusieurs changement(s) d'état à deux instants successifs de temps ou sur un intervalle de temps (ex. entre, sort).

Un **événement primitif** est un changement d'état.

Un **événement composé** est une combinaison d'états et/ou d'événements.

Un même événement peut être regardé à différentes granularités spatiales et temporelles. Par exemple, un homme qui court dans un Marathon peut être vu comme un état (il court) ou comme événement composé. La granularité choisie indique les propriétés d'intérêt pour l'utilisateur.

3.3. Description de la hiérarchie de notre ontologie

La création de notre ontologie est basée sur une syntaxe formelle qui inclue la plupart des concepts pouvant appartenir au domaine de la vidéo surveillance. En effet, notre approche représente une extension du travail réalisé par San Miguel et al. [44]. Cette extension est dû au fait d'insertion de nouveaux concepts utilisés dans le domaine de l'industrie, vu qu'il existe un intérêt immense par rapport au test de l'analyse du contenu vidéo dans le domaine de la vidéo surveillance comme par exemple l'évènement de l'intrusion.

Généralement, l'interprétation sémantique des séquences vidéo représente l'étape critique dans le processus d'indexation. Cette interprétation correspond à la traduction des données issues du module d'analyse du bas niveau en un sens sémantique. Ici, on utilise le paradigme d'ontologie comme un moyen avec lequel on caractérise le système de vidéosurveillance. Afin d'atteindre ce but, notre ontologie sémantique est représentée par différents concepts qui sont en interaction entre eux. De plus, chaque concept comporte une ou plusieurs propriétés comme Data_Property, qui sont présentées en détail dans ce qui suit.

3.3.1. Les concepts de notre ontologie

Les concepts de l'ontologie présentée dans ce travail de thèse correspondent à la catégorisation du domaine de la vidéo surveillance, considéré comme des relations de généralisation/spécialisation. Dans le but d'avoir une représentation complète de tous les objets et les évènements qui peuvent surgir dans le domaine de la vidéo surveillance, nous avons divisé

formellement notre ontologie en quatre grandes catégories de concepts, représentant **Video_Actions, Video_Events, Video_Objects et Video_Sequences**.

La Figure 12 ci-dessous, démontre la liaison entre les quatre catégories de concepts, ou chaque concept forme une interconnexion avec les autres. Premièrement, toutes les Vidéo séquences existantes dans la base de données vidéo doivent être indexées avec un ou plusieurs concepts qui appartiennent à la catégorie Video_Events. Une relation existe entre les catégories Video_Events et Video_Actions. Plus précisément, on considère un évènement comme une composition d'une ou plusieurs actions. De plus, une autre connexion existe entre les catégories Video_Actions et Video_Objects. En effet, la description de scène est formée par le composant de la catégorie Video_Object où réside l'action. Finalement, les séquences vidéo appartenant à la catégorie Video_Sequences regroupent les objets de la catégorie Video_Objects.

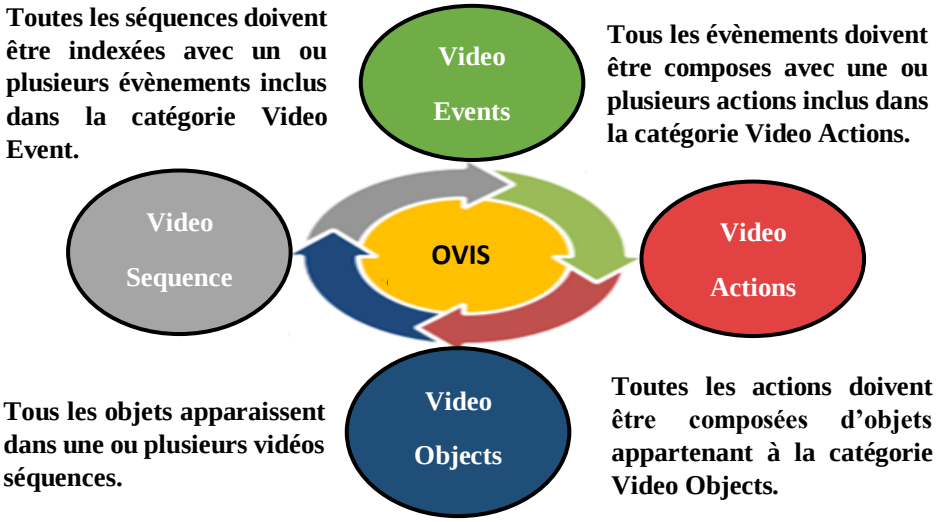


Figure 12. L'interconnexion entre les quatre grandes catégories de concepts de notre ontologie de vidéosurveillance.

a. La catégorie Vidéo_Actions

La catégorie Video_Actions inclue les actions qui peuvent être repérées dans les évènements de la vidéo surveillance. Cependant, les différentes variétés d'objets peuvent produire plusieurs types d'actions. Généralement, on peut trouver cinq catégories d'objets : **Airplane, Boat, Train, Traffic (Road Traffic)** et **humans**. Dans notre ontologie, nous avons divisé ces actions sur plusieurs sous classes selon la nature des objets qui sont en interaction :

- Interactions avec l'environnement: comme human_walking, human_stopping, airplane landing, etc…

- Interactions avec les êtres humains : comme human_attacking, human_meeting, etc...
- Interactions avec les objets : comme human_breaking_an_object, etc…

Dans le tableau 2 ci-dessous, nous avons résumé toutes les sous-classes de la partie Video Actions de notre ontologie, illustrées à plusieurs niveaux (N1 a N3). Comme décrit, nous avons séparé toutes les actions selon leurs degrés de priorité. Le premier degré de sous classes exprimé dans le deuxième niveau représente les acteurs de l'action les propriétés avec qui elle interagit (e.g, **Human_And_Objects_Actions**...). Le second degré de sous classes exprimé au niveau 3 correspond aux actions eux-mêmes, parmi elles, on peut citer celles qui sont reportées depuis la littérature come (**Split, Met, Ran**) et celles relatives au domaine industriel come (**Crossed_Virtual-Line, Trespassed**).

N 1	N 2	N 3
Video_Actions	Train_And_Environement_Actions	Fire_detected/ Arrival_detected/ Stationed
	Train_And_Human_Actions	Upped/ Downed/ Crossed_Forbidden_Zone
	Boat_And_Environement_Actions	Navigated
	Boat _And_Human_Actions	Threw_bag
	Boat _And_Objects_Actions	Approached_bank/ Drived_away_the_bank
	Airplane_And_Environement_Actions	Crashed_Off / Flew / Landed / Took_Off
	Airplane_And_Human_Actions	Disembarked / Embarked
	Airplane_And_Objects_Actions	Registered_Luggage / Took_Luggage
	Human_And_Objects_Actions	Downed_Stairs / Read _Document / Broke_Object / Browsed_Object /Counted_Ground-Vehicule /Counted_Human / Crossed_Virtual-Line/ Left_Luggage / Put_Object /Removed_Luggage / Rested_On_Chair / Smoked_Cigarette / Upped_Stairs
	Human_And_Human_Actions	Attacked / Chased / Evacuated /Fell_Down / Flow_Opposed / Formed /Fought / Helped / Hit /Local_Dispersed / Met / Split / Stole /Talked / Waited
	Human_And_Environement_Actions	Appeared / Counted_Speed /Disappeared / Entered_Area / Face_Recognized / Fell / Intruded /Left_Area / Loitered / Overcrowded /Ran / Skateboarded / Slipped /Stopped / Trespassed / Walked
	Traffic_And_Object_Actions	Crashed Object
	Traffic_And_Environement_Actions	Parked

Tableau 2. Représentation hiérarchique de la partie Video_Actions.

b. La catégorie Video Events

Durant les dernières années, une intention particulaire a été attribuée dans l'état de l'art dans le domaine de la détection d'action qui correspond à l'aspect le plus important dans le domaine de la vidéo surveillance. Dans la présente ontologie, la catégorie Video_Events comprend tous les différents évènements qui peuvent se produire dans le flux de la vidéo surveillance. Chaque évènement représente la formation d'actions comprenant un ou plusieurs objets d'intérêt qui sont en interaction.

Le tableau 3 ci-dessous décrit les quatre niveaux (N1 a N4) représentant les différents sous classes de la partie Video_Events de notre ontologie. Chaque niveau correspond à un degré de priorité, comme décrit dans la Section 3.2. Le premier degré est relatif au nombre d'objets d'intérêt et divise la catégorie Video_Events en deux grandes sous classes représentant respectivement les évènements d'objets individuels et multiples. Le second degré est relatif à la nature d'objets représenté par 7 types : **Group-Of-Person, Multiple_Ground-Vehicle, Airplane, Train, Boat, Person and Singles_Ground-Vehicle**. Le degré final correspond à l'interaction entre ces objets et les propriétés avec lesquelles elles interagissent comme **Humans, Environment** ou **Objects**.

L 1	L 2	L 3	L 4
Video_Events	Multiple_Objects _Events	Group-Of-Person_Events	Interaction_Group-Of-Person_And_Environement / Interaction_Group-Of-Person_And_Human
		Multiple_Ground-Vehicle_ Events	Interaction_Multiple_Groundvehicle_And_Objects
	Single_Objects_E vents	Air-Plane_Events	Interaction_Airplane_And_Environement/ Interaction_Airplane_And_Human / Interaction_Airplane_And Objects
		Train_Events	Interaction_Train_And_Environement/ Interaction_Train_And_Human
		Boat_Events	Interaction_Boat_And_Environement/ Interaction_Boat_And_Human / Interaction_Boat _And_Objects
		Person_Events	Interaction_Person_And_Environement / Interaction_Person_And_Human/ Interaction_Person_And_Objects
		Single_Groud-Vehicle_Events	Interaction_Single_Ground-Vehicule_And_Environement

Tableau 3. Représentation hiérarchique de la partie Video_Events.

c. La catégorie Vidéo Objects

Les différents types d'objets représentent les entités principales qui interagissent dans les séquences vidéo. La catégorie Video_Objects inclue tous les objets qui peuvent apparaitre dans la scène de vidéo surveillance. Une grande variété d'objet en interaction entre eux crée les actions de vidéo surveillance. Selon l'aspect de leur mobilité, les objets peuvent être regroupés dans deux grandes catégories. Les objets contextuels qui n'ont aucun aspect de mobilité et les objets mobiles qui ont cet aspect de mobilité. Eventuellement, on peut ajouter une troisième catégorie représentant les caractéristiques d'image (ROI), sous le nom de **Low_Level_Features**. Elles représentent en réalité toutes les données extraites depuis le module d'analyse vidéo. Le tableau 4 ci-dessous illustre ses trois catégories et divise la catégorie Video Objects en cinq niveaux (N1 aN5).

La catégorie d'objet contextuel est divisée en deux sous classes représentant les objets fixes (qui ne peuvent jamais être bougés) et les objets Portables (qui ont la possibilité d'être bougé). Les objets fixes correspondent à ceux de la création humaine (e.g. **Air_Conditioners, Panels**, etc.) et les objets naturels (e.g. **Grass, Land**, etc.). La partie d'objets portables représente tous les objets avec aspect d'utilisation générale ou particulière (e.g. **a Box, a Chair, a CellPhone**, etc.).

La catégorie d'objets mobiles inclue quatre sous classes qui ont la possibilité d'auto-bouger comme (**Animals, Airplanes, Trains, Boat, Human and Traffic**). Ici, le mot Traffic correspond à toutes les Bicyclettes et Mobilettes. La sous classe Human correspond à l'entité humaine comme (**Person** ou **Group-Of-Person**).

La catégorie Low-Level Features regroupe tous les résultats du module d'analyse bas niveau qui peuvent être utilisés pour aider le processus d'indexation pour le but principal de la détection d'évènement, comme (**Bounding Box, Frame**, etc.)

d. La catégorie Video_Sequences

La catégorie Video-Sequences représente la classe de toutes les vidéos indexées par notre system OVIS et ses instances représentant la base de données vidéo.

L 1	L 2	L 3	L 4	L 5
Video_ Objects	Contextual_Objects	Fixed_Objects	Human_Creation _Objects	Air-Conditioner /Building /Electrical-Pole / Equipment / Floor /Panel / Parking-Lot / Road / Stairs /Stairs-Barrier / Wall /Zone / Glass Barrier
			Natural_Objects	
		Portable_Objects	General_Using	Box / Chair / Door /Plant /Reception-Desk / Surveillance-Camera /Table / Window /Curtain / Sofa
			Self_Using	CellPhone /Document / Luggage(Bag ; Suitcase)
	Mobile_Objects	Airplane		
		Boat		
		Train	City-Tramway / Long-Distance-Train / Underground	
		Animals		
		Human	Person / Group-Of-Person	
		Ground_Traffic	Bicycle	
			Ground-Vehicle	Bus / Car / Truck
			Motorcycle	
	Low_Level_features	Bounding-Box		
		Frame		
		Major_BoundIng-Box		
		Temporary_ Bounding-Box		
		Temporary_Group-Of-Person		

		Blocs		

Tableau 4. Représentation hiérarchique de la partie Video_Objects.

3.3.2. Les DataProperty de notre ontologie

Les **Data_Property** représentent l'information réelle relative aux concepts des individus. Dans notre ontologie, Data_Property inclue toutes les propriétés relatives à un ou plusieurs concepts. Le tableau 5 ci-dessous présente la hiérarchie des Data_Property divisée en trois niveaux (N1 a N3). Le niveau le plus haut est divisé quant à lui en sept sous-classes relatives au type de la Data_Property, comme les propriétés des évènements, les propriétés des images, etc. Chacune d'elle inclue une ou plusieurs Data_Property (**Event_Place**, **Video_URI**, etc.)

3.3.3. Les Object_Property de notre ontologie

Les **Object_Property** concernent les interactions entre les concepts de notre ontologie et sont divisées en trois niveaux (N1 a N3) comme présenté dans le tableau 6. Afin d'avoir une représentation complète de toutes les interactions entre les concepts de notre ontologie, on a subdivisé le niveau le plus haut en trois sous classes qui reflètent les objets en interactions. On considère deux catégories d'objets comme Humans et Objects ; tandis que, les interactions représentent **Human_Against_Human, Human_Against_Objects**, et **Objects_Against_Objects**. Dans le dernier niveau, chaque type d'interaction englobe ses propriétés d'objets, comme **Asked_Direction, Walked_Around, Detected_In**, etc.

L 1	L 2	L 3
Top_Data_Property	Event_Properties	Event_Place / Nature_Event
	Detected_Objects	Bottom_Left_Point_X / Bottom_Right_Point_Y / Detected_In_Frame / Direction/ End_Frame / Height / ID / Leaving_Object_Way / Major_BB / MBB_True / Number / Number_Of_Person / Posture / Speed / Start_Frame / Top_Left_Point_X / Top_Right_Point_Y / Weight ...
	Entering_Exit	
	Frame_Properties	Number_Frame /Number_BB_In_Frame /Number_MBB_In_Frame /Started_MBB ...
	Type	
	Time	
	Video_Sequence_Properties	Video_URI / Number_Of_Frame / Start_Frame_Event / End_Frame_Event …

Tableau 5. Représentation hiérarchique de la partie Top_Data_Property.

L 1	L 2	L 3
Top_Object_Property	Human_Against_Human	Asked_Direction / Chased /Formed_Final_Meta_Group / Attacked / Formed_With / Had_Diff_End_Position / Had_Different_Direction / Had_Same_Start_Position / Had_Started_Meta_Group / Helped / Hit / Met_With / Pushed / Split_With / Spoke_With / Stole / Walked_With ...

	Human_Against_Objects	Walked_Around / Attempted_To_Open /Browsed_On / Downed / Left / Loitered_In / Occurred_In / Put / Rested_On / Stood_Near / Upped ...
	Objects_Against_Objects	Detected_In / Crashed_With / Flew_In / Landed_In / Parked_In / Represented /Took_Off_From ...

Tableau 6. Représentation hiérarchique de la partie Top_Object_Property.

3.3.4. Convention de nommage de notre ontologie

Afin d'obtenir une ontologie formelle et cohérente, on a créé une convention de nommage composé de plusieurs concepts.

Premièrement, tous les nouveaux concepts créés dans la catégorie Video_Event doivent être nommés de la même façon comme les précédentes de cette catégorie. Chaque concept est composé de trois parties :

- Les interactions (chaque concept débute avec le nom d'interaction).
- Le nom d'objet existant dans la catégorie Video Object.
- La propriété avec laquelle cet objet interagit.

Ensuite, tous les concepts de notre ontologie sont généralisés et leurs détails sont classifiés sous la forme d'Object_Properties et Data_Properties. Comme exemple typique, la notion du temps, de la posture, de la position et d'interaction sont classés comme Object_Properties ou Data_Properties, et non comme sous classe d'évènement.

La duplication dans le nommage des concepts doit être évitée et les évènements multiples séparés des évènements simples, aussi les actions des humains de celles des objets, tout comme les événements de personnes de ceux des objets.

De plus, les Data_Properties et les Object_Properties doivent être généralisés et les duplications évitées. Par exemple, à l'inverse d'avoir une Data_Property nommée **Name-Of-Animals** du concept **Animal** et une autre nommée **Name-Of-Building** du concept **building**, on introduit une généralisation et on nomme la Data_Property "**Name**" afin de l'utiliser pour les deux concepts cités auparavant.

Les événements et les actions doivent être séparés selon leur degré de priorité. Concernant les évènements, le premier degré est attribué au nombre d'objet (multiple ou simple). Le second degré est lié à la nature de l'objet (**Group-Of-Person, Person, Ground-Vehicle**, etc.), et le troisième représente l'interaction d'objet (avec les humains, avec l'environnent ou avec les objets). Le premier degré est attribué selon les acteurs des actions et les propriétés avec lesquelles elles interagissent, tandis que le second degré concerne l'action elle-même.

Dans notre travail, Group-Of-Person est considéré comme une formation de deux ou plusieurs personnes.

Pour avoir plus de précisions, chaque mot dans la formation des concepts, des Object_Property ou des Data_Property, doit débuter avec une majuscule. Une relation est déterminée entre les concepts (Object_Property) dans les trois catégories : **Object_Against_Object, Human_Against_Object, Human_Against_Human**.

Les nouveaux concepts ajoutés à la catégorie Video_Actions sont représentés en un seul mot définissant l'action, à l'exception de la catégorie Object, où nous avons besoin de spécifier les interactions de nos concepts (**Human, Ground-Vehicle, Airplane**).

La nature de l'interaction avec l'objet dans la catégorie Video_Actions doit être spécifiée si l'action ne considère pas tous les types d'objet.

En plus, tous les concepts de notre ontologie sont créés en unifiant les mots avec le symbole (_), à l'exception des objets composés qui sont liés avec le symbole (-).

3.4. Domaines d'applications de notre ontologie de vidéosurveillance

Dans notre travail de thèse, on a proposé une ontologie formelle et complète qui couvre la majorité des évènements d'une vidéosurveillance. Cette représentation de connaissance complète peut être utile dans plusieurs domaines d'applications comme par exemple :

3.4.1. Création des benchmarks

Les benchmarks représentent des challenges de validation qui sont très utilisés par la communauté des chercheurs dans le domaine de la vision par ordinateur. Par exemple, la vidéosurveillance utilise les benchmarks comme PETS, TRECVID, CAVIAR, etc. pour la reconnaissance et la détection d'évènements. On tient à rappeler que notre ontologie de vidéosurveillance proposée dans cette thèse permet d'appuyer ces benchmarks dans le processus de sélection des évènements appropriés.

3.4.2. Description des scènes

La description d'une scène représente tous les objets qui sont joués dans une vidéo. Ces derniers forment l'arrière-plan d'une vidéo (i.e. les objets qui apparaissent longtemps dans la scène durant une longue durée) et/ou l'avant plan d'une vidéo (i.e. les nouveaux objets qui viennent de s'introduire dans la scène).

Récemment, une attention particulière est attribuée au processus de description automatique d'images. Kuznetsova et al. [68] proposent une approche d'annotation manuelle de description d'images à une large échelle qui est basée sur la recherche similaire des images capturées depuis une large base de données, avant de générer de nouvelles descriptions par la généralisation et la recomposition des captures recherchées. Cette approche inclue typiquement une étape de généralisation intermédiaire pour la suppression des spécificités des captions qui sont optimales uniquement pour les images recherchées comme le nom d'une ville par exemple. L'approche proposée par Socher et al. [69] utilise une représentation indépendante et un réseau de neurones pour regrouper des images avec des phrases sous forme d'un vecteur commun. Cette approche montre comment tracer la représentation de phrases depuis un réseau récursif dans le même espace réservé aux images. Vinyals et al. [70] démontrent l'efficacité de stockage du contexte d'information dans une couche récurrente, et développent une approche générative basée sur une combinaison de réseaux de neurones récurrente pour générer des captures d'images en traitant leurs résultats sur des propriétés d'images extraites par un réseau de neurones. Cette approche utilise la base de données MS COCO qui contient actuellement 5000 images avec 40 références de phrases pour traiter la précision des mesures automatiques. Kiros et al. [71] utilisent deux chemins différents (pour l'image et le texte) pour définir leur regroupement, bien qu'ils puissent produire le texte. Ils proposent une architecture différente en utilisant un état caché de l'encodeur LSTM (Long Short-Term Memory) au temps « T » comme une représentation de séquence d'entrée codée de longueur « T ». Ils tracent cette représentation de séquence et la combinent avec la représentation visuelle du modèle Convnet. Un espace joint est obtenu depuis un décodeur de prédictions de mots.

Figure 13. Description de scènes des benchmark PETS 2004 et PETS 2012.

Mao et al. [72] considèrent de nouvelles perspectives pour les méthodes bidirectionnelles qui recherchent les images basées sur des entrées textuelles, ou les phrases depuis l'image donnée. Ils développent des méthodes robustes d'apprentissage commun depuis les entrées d'image et de texte pour former un très haut niveau de représentation depuis les modèles comme le CNN (Convolutional Neural Networks). Ils testent leurs méthodes sur la reconnaissance d'objets et de mots précis depuis un corpus de texte à large échelle. Ils proposent un système utilisant un réseau de neurones convolutionnel pour extraire les caractéristiques d'images, et un autre réseau de neurones récurrent pour les phrases, avec une interaction réalisée dans une couche multi-modèle commune. La Figure 13 ci-dessus montre une segmentation idéale et très précise de deux scènes extraites depuis les benchmark PETS [73]. La scène contient des éléments statiques qui ne changent pas à travers le temps (i.e. buildings, grass, electric poles, roads, trees, car parks, restrictive roads). Ces concepts ont été utilisés dans notre ontologie de vidéosurveillance.

3.4.3. Description des vidéos

Comme pour le processus de description d'images, celui des vidéos représente une opération décrivant les objets animés. Cependant, la différence réside dans le fait que les images sont statiques tandis que les vidéos nécessitent des informations qui ont un rapport avec la notion de temp.

Plusieurs travaux de la littérature se sont intéressés à la description des vidéos. Yao et al. [74] développent des modèles détaillés qui utilisent la description (Long-Short-Term Memory "LSTM"), permettant de sélectionner les segments temporels les plus adéquats dans les vidéos et d'incorporer un réseau CNN 3D en générant des phrases. Deux types d'encodeurs sont testés : le premier est une application d'encadrement d'images, tandis que le second est un réseau 3D convulotionnel. Rohrbach et al. [75, 76] utilisent une approche basée sur une translation d'une machine qui produit des descriptions de vidéo de plusieurs individus qui cuisinent dans le même endroit, en partant d'une représentation sémantique intermédiaire jusqu'à la génération de phrases. Ces phrases sont générées à partir d'une représentation de rôle sémantique des concepts de haut niveau comme des acteurs, des actions et des objets. Venugopalan et al. [77] appliquent l'approche des neurones à la génération statique des captures d'images et utilisent le décodeur LSTM pour la tâche de génération des descriptions automatiques des vidéos. Ils utilisent un

réseau de neurones convolutionnel pour l'extraction des caractéristiques d'apparence depuis chaque image d'un clip vidéo.

3.4.4. Indexation des évènements vidéo

Le problème d'indexation et de recherche des vidéos est considéré comme un axe de recherche d'un intérêt très remarquable dans le domaine de la vision par ordinateur à cause de la taille grandissante des bases de données générée depuis les enregistrements vidéo. En effet, nous avons proposé un nouveau système d'indexation OVIS [6] qui utilise une ontologie pour la détection d'évènements dans une vidéosurveillance.

3.5. Conclusion

Dans le domaine de la vidéosurveillance, l'ontologie a prouvé son efficacité du fait qu'elle peut être utilisée dans le processus d'indexation pour la détection d'évènements. Dans ce chapitre, nous avons présenté une ontologie de vidéosurveillance en expliquant sa syntaxe et sa sémantique. Notre ontologie décrit tous les objets et les évènements qui peuvent exister dans le domaine de la vidéosurveillance et elle a été élargie aux nouveaux concepts qui caractérisent les évènements dans les vidéos industrielles. Un autre point positif de notre approche est que son utilisation concerne plusieurs domaines d'application, à savoir : l'indexation et la recherche des vidéos, la création des benchmarks, la description de scènes, etc...

Dans le chapitre suivant, on va implémenter et mettre en œuvre notre ontologie pour l'indexation et la recherche d'évènements dans une vidéosurveillance, en expérimentant notre système OVIS avec les vidéos des [1]benchmark **PETS 2012**[1] et **TRECVID 2016**[2],

[1] http://www.cvg.reading.ac.uk/PETS2012/a.html

[2] http://www nlpir.nist.gov/projects/tv2016/tv2016.html

Chapitre IV : Implémentation, Expérimentation et Comparaison de notre Prototype (OVIS)

4.1. Introduction

Les avancées technologiques réalisées durant cette dernière décennie dans le domaine de l'informatique (espaces de stockage de plus en plus considérables, caméras FHD, numérisation des données, etc.) ont permis de simplifier l'utilisation de données vidéo sur le Web et dans le domaine de la vidéosurveillance. En effet, l'indexation et la recherche multimédia ont vu le jour et ont fait l'objet de plusieurs travaux de recherche sur l'analyse du comportement humain. Le but est de définir la sémantique du comportement humain qui est considéré comme une information de bas niveau en utilisant des méthodes d'extraction d'information de bas niveau tel que les blobs. Pour cela, une ontologie de vidéosurveillance est utilisée pour résoudre le problème de la sémantique relative à cette séquence vidéo et de déduire ainsi le comportement correspondant.

Dans ce chapitre, nous allons concevoir et réaliser un système d'indexation et de recherche des vidéos. Nous allons présenter tout d'abord la modélisation de l'architecture de notre système ainsi que l'environnement matériel et logiciel utilisé pour sa mise en œuvre. Ensuite, nous allons décrire en détail l'implémentation d'un prototype du système avec les résultats obtenus en expérimentant plusieurs vidéos des benchemarks PETS 2012 et TRECVID 2016. Une étude comparative de notre solution proposée avec celles de la littérature montrera sa consistance et son efficacité dans l'indexation et la recherche des vidéos. Enfin une conclusion sera donnée à la fin du chapitre.

4.2. Modélisation UML du prototype

Afin de représenter les fonctionnalités attendues de notre système proposé, nous avons opté pour le langage UML (*Unified Modeling Language*) [78] qui est un langage de modélisation permettant d'exprimer des modèles objet et couvre toutes les phases du cycle de vie d'un logiciel. UML n'étant pas une méthode puisqu'elle n'offre pas une démarche à suivre. Cependant, UML est un langage formel de communication, qui facilite la représentation et la compréhension d'une solution objet. Dans notre cas, nous avons utilisé trois diagrammes UML pour modéliser un prototype de notre système à savoir : le diagramme de cas d'utilisation, le diagramme de séquence et le diagramme de déploiement.

4.2.1. Diagramme de cas d'utilisation du système

Le diagramme de cas d'utilisation sert à décrire l'expression des besoins et se présente sous deux formes (un schéma regroupant les activités d'un acteur et un tableau pour détailler les actions liées à une activité). Au niveau de notre système d'indexation OVIS, le diagramme de cas d'utilisation nous permet de modéliser les principales fonctions dont l'utilisateur du système doit disposer. Comme l'illustre la Figure 14, l'utilisateur du système a besoin de deux cas d'utilisation principale à savoir : l'indexation et la recherche des vidéos. Cependant, ces deux cas d'utilisation engendrent d'autres cas à travers des relations « include » et « extend ».

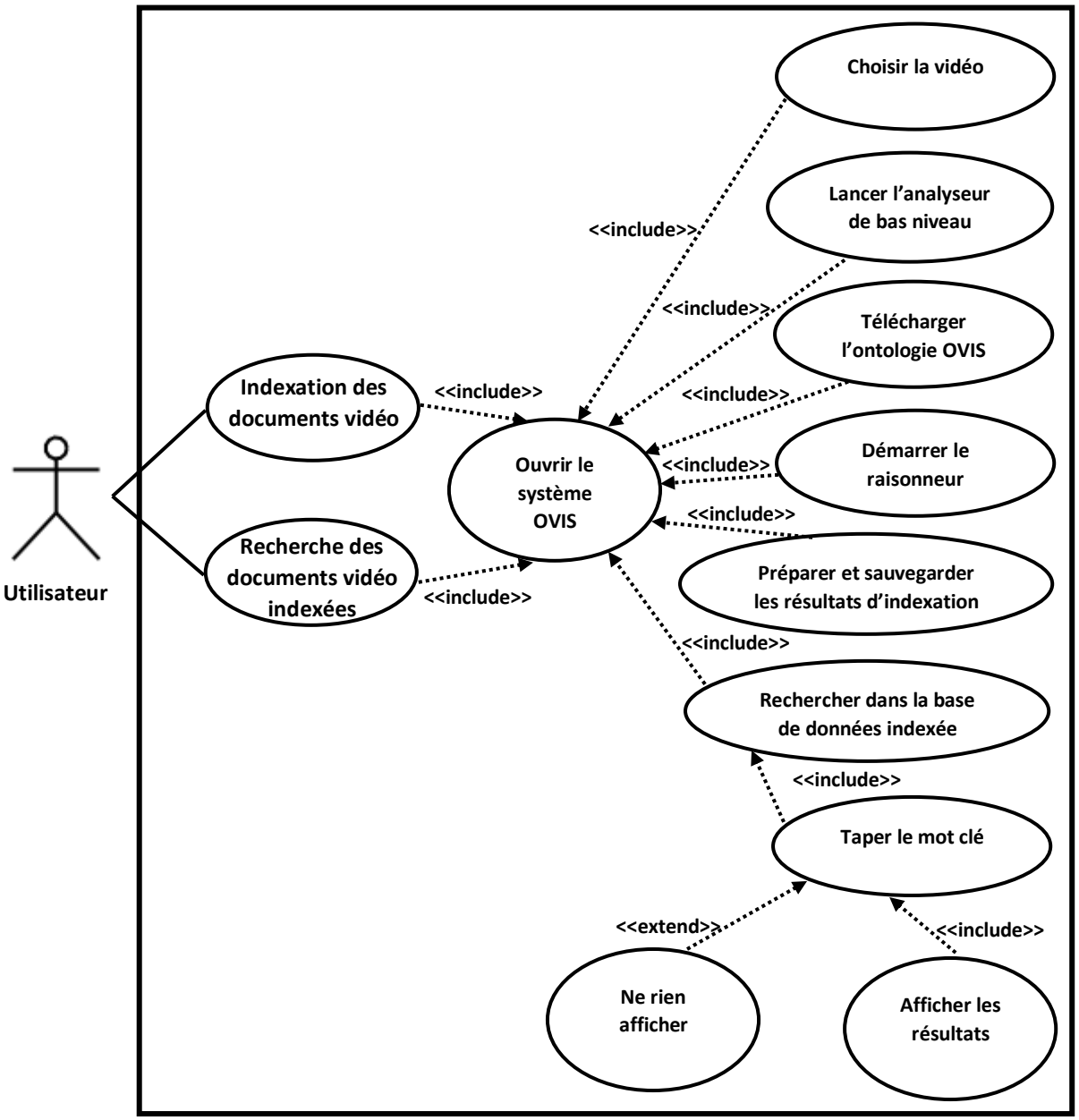


Figure 14. Diagramme de cas d'utilisation générale.

4.2.2. Diagramme de séquence du système

Le diagramme de séquence représente la schématisation de messages entre les différents objets d'un système informatique afin de réaliser un cas d'utilisation. Ces messages représentent généralement des appels de méthodes qui permettent de faire interagir les différents objets. Deux cas d'utilisation sont représenté (le cas d'utilisations d'indexation et celui de recherche des documents vidéo). Comme l'illustre la figure 15, le premier diagramme de séquence représente le cas d'utilisation « indexation des documents vidéo ». L'utilisateur démarre le système et commence par choisir la vidéo qu'il veut indexer. Il va ensuite exécuter l'analyseur de bas niveau afin d'extraire les informations physiques lui permettant de mettre à jour le fichier OWL de l'ontologie (ajout des Individuals, Data-Properties et Object-Properties). Ensuite, il va démarrer le raisonneur Pellet qui va exécuter les différentes règles d'inférence SWRL afin d'extraire les évènements de la séquence vidéo. Enfin, cette séquence vidéo indexée sera stockée donc dans une BDD avec ses métadonnées.

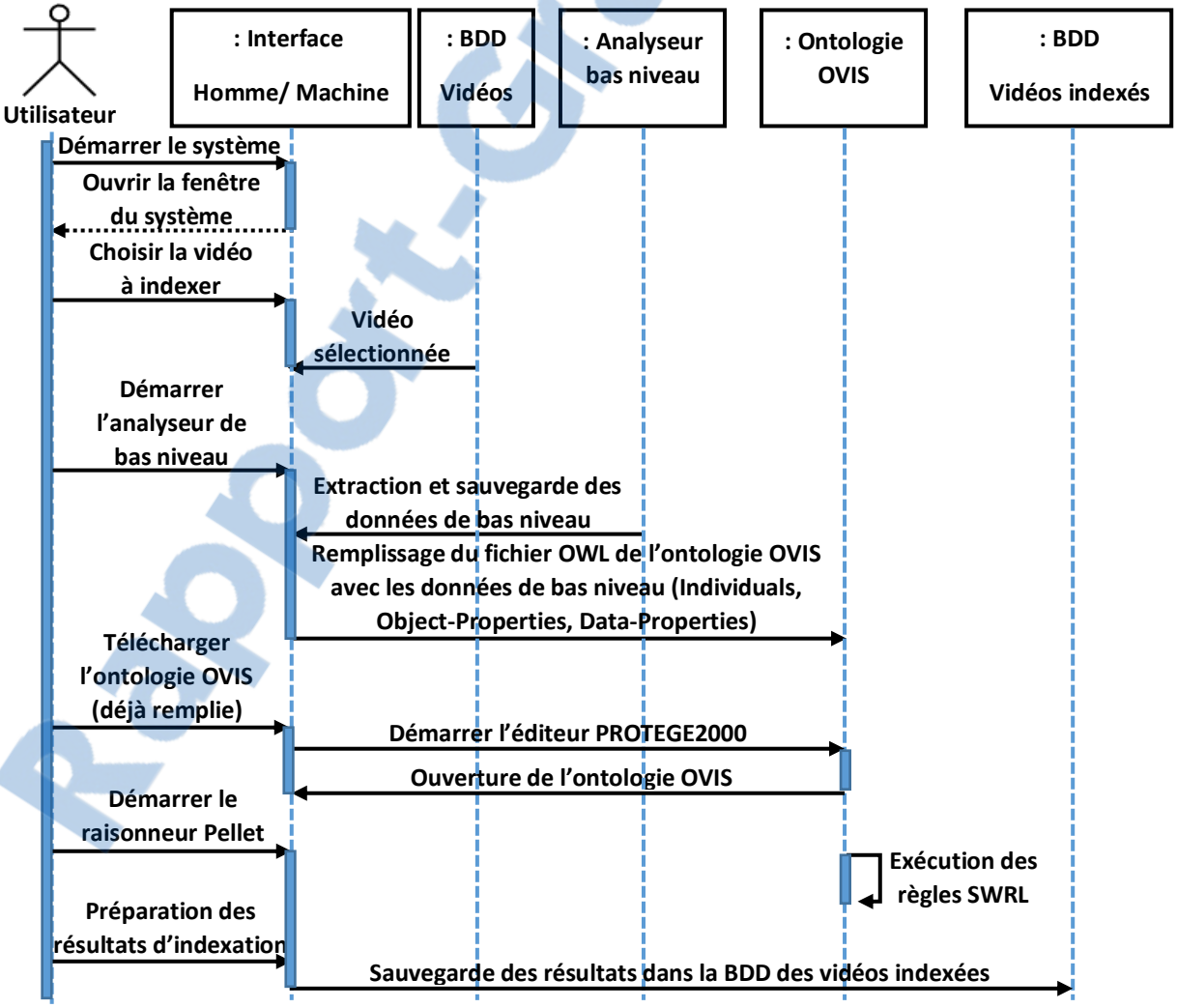


Figure 15. Diagramme de séquence du cas « Indexation des documents vidéo ».

La figure 16 illustre le deuxième diagramme de séquence « recherche des documents vidéo ». En effet, l'utilisateur démarre le système afin de faire une recherche dans la BDD des séquences vidéo indexées. Il doit saisir le mot clé (un évènement) qu'il souhaite rechercher et le système lui renvoie alors le résultat de sa recherche en affichant les séquences vidéo indexées, ainsi que le frame d'image de début et celle de la fin de l'évènement dans chaque séquence vidéo.

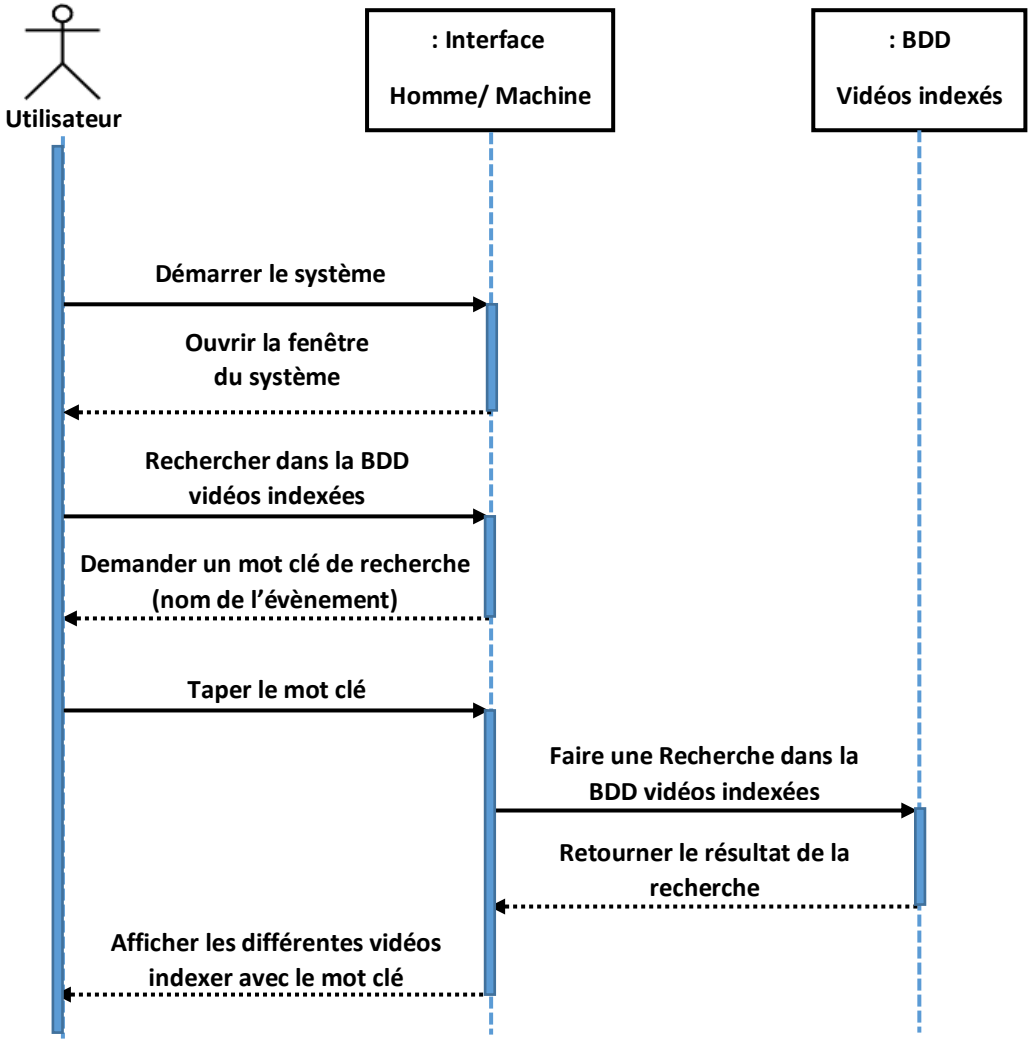


Figure 16. Diagramme de séquence du cas « Recherche des documents vidéo indexées ».

4.2.3. Diagramme de déploiement du système

Le diagramme de déploiement représente l'organisation physique de la distribution des composants logiciels de l'application. Ce type de diagramme engendre trois ensembles d'éléments (composants, nœuds et associations). En effet, les nœuds sont des éléments physiques qui représentent des ressources matérielles informatiques et qui sont reliés entre eux par des associations physiques (câblage etc..). Cependant, les composants représentent les

éléments logiciels de l'application. La figure 17 illustre les nœuds matériels et les composants logiciels de notre système OVIS. Les différents composants logiciels (BDD vidéo, Ontologie, Analyseur de bas niveau et la BDD Vidéos Indexées) communiquent avec le composant principal (Système OVIS), sous le nœud physique (Ordinateur Portable).

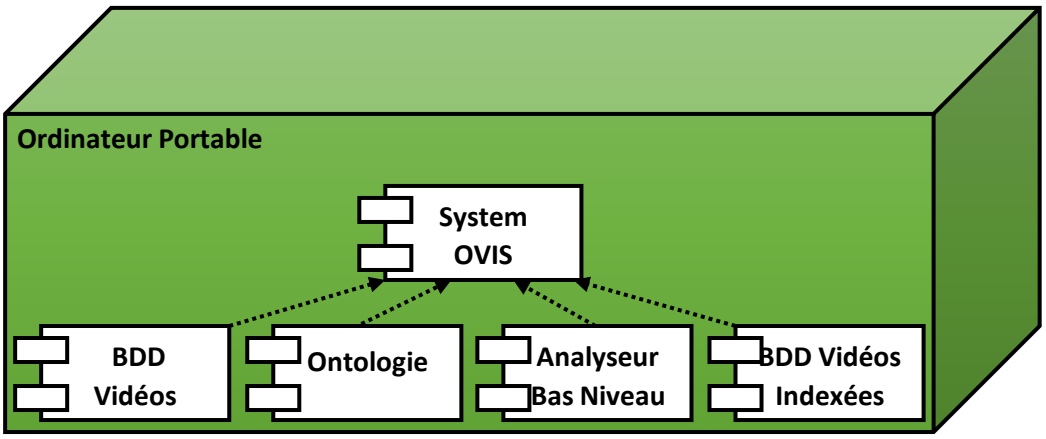


Figure 17. Diagramme de déploiement du système.

4.3. Conception et implémentation de notre prototype

Dans OVIS, l'ontologie représente le noyau de l'architecture globale du système comme l'illustre la Figure 18 ci-dessous. Elle permet d'assurer le processus d'indexation et de recherche des vidéos depuis l'étape de détection des caractéristiques des blobs dans les Bounding Box par le module d'extraction jusqu'à la dernière étape qui représente l'identification et l'indexation des évènements dans une séquence vidéo.

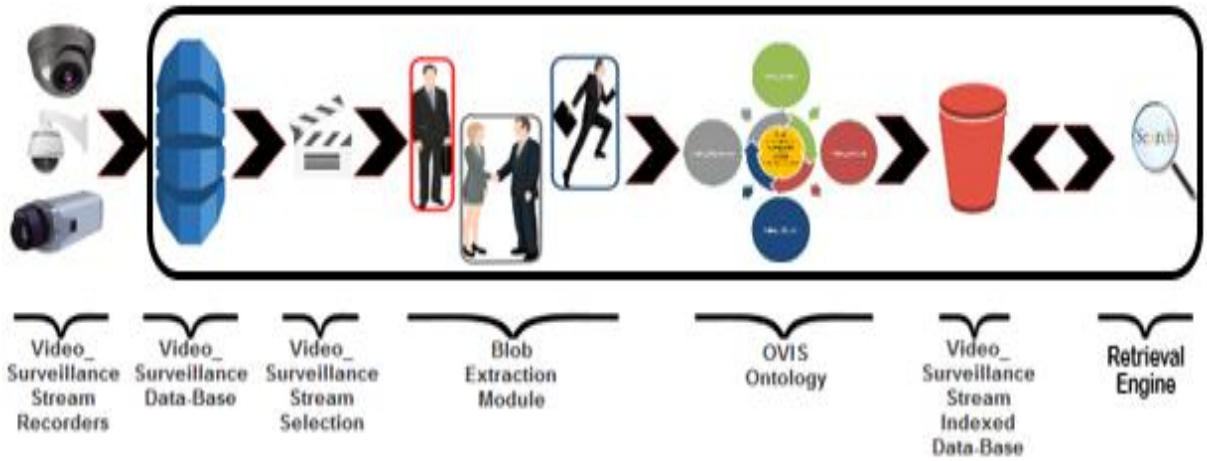


Figure 18. Architecture générale de notre system OVIS.

Le processus d'indexation vidéo illustré dans la Figure 18 débute lorsque le module d'analyse vidéo va extraire les différentes caractéristiques de bas niveau des blobs bounding box de la séquence vidéo en utilisant la méthode du "Background Subtraction" comme « Top Left Point X, Top Left Point Y, Width, and Length ». L'ontologie utilise ces caractéristiques pour les organiser afin de créer les « Data Property » et les « Object Property » etc. Le raisonneur de notre ontologie les ordonne ensuite correctement en utilisant un ensemble de règles SWRL et indexe cette séquence vidéo dans la classe Video_Action selon le comportement de ses objets. Finalement, la vidéo va être indexée et sauvegardée dans une base de données pour des utilisations futures. Pour le processus de recherche, notre system OVIS permet la recherche de toutes les séquences vidéo indexées dans la base de données en utilisant des mots clés exprimés sous forme de noms d'actions (Walking, Running, Splitting, and Formation). Par exemple, si on veut rechercher l'évènement Walking, on utilise Walking comme mot clé et notre system OVIS retourne alors toutes les séquences indexées avec ce mot clé.

4.3.1. Le module d'extraction des blobs

La meilleure approche de détection d'objets est connue sous le nom de « Blob Regions » à cause du nombre élevé de domaines quelle couvre. Cette caractéristique laisse les « Blobs » comme meilleure représentation que les « Points, Corners ou Edges ».

Pour cela, différents algorithmes existent pour la collecte des blobs. La méthode de la soustraction de l'arrière-plan (Background Substraction) permet une classification des pixels de l'image en avant-plan et arrière-plan. Les différents blobs sont extraits en collectant les pixels de l'avant plan de l'image. La méthode du flux optique peut être utilisée pour l'extraction des caractéristiques de chaque pixel dans chaque mouvement d'image. Ses flux sont alors groupés dans des blobs avec des mouvements cohérents et sont traités par une matrice Gaussienne multi-variante. Les flux optiques sont utiles pour caractériser chaque mouvement de pixels selon certaines caractéristiques du vecteur de flux.

Dans notre travail, la méthode du background soustraction est utilisée pour extraire les caractéristiques des blobs qui peuvent apparaitre dans chaque image avec leurs Bounding Boxes. Ces caractéristiques représentent les données de bases de notre approche ontologique à base de règles SWRL pour la détection des évènements (voir l'algorithme ci-dessous).

Blob extraction features Algorithm

```
while(true) do
    cap.init(img);
    if ( img.empty() ) break; // stop if end of processing
    if (nframes = 0) then // preparing the first frame
                    labelImg ← CreateLBLImage(img);
                    BlobsImg ← CreateBlobImage(img);
                    Clear(BlobsImg);
                    morphKernel ← InitKernel();
    end if
    if ( fgimg.empty() ) then
                        Init(fgimg,img);
    end if
    //update the model
    bg_model.Update(img, fgmask)
    fgmask.InitThreshold();
    fgmask.InitmorphologyEx(morphKernel);
    fgmask.dilate();
    fgimg.ScalarAll(0);
    img.copyTo(fgimg, fgmask);
    BlbDetectionResult ← getBlob(fgmask, labelImg, blobs);
    blobs.FilterByArea();
    FilterLabels (labelImg, BlobsImg, blobs);
    vector contours;
    contours.findContours(BlobsImg);
    contours.drawContours(img);
    outfile ← nframes;
    if (blobs.size()>0) then
                        compteur++;
    end if
    int compteurblb ←1;
    for all blobs do
                    blbRect.getMinx(blobs);
                    blbRect.getMiny(blobs);
                    blbRect.getHeight(blobs);
                    blbRect.getWidth(blobs);
                    Mat blobMask;
                    blobMask.setMask(blbRect);
                    long size ← countNonZero(blobMask);
                    int w ← blobs.getMaxX - blobs.getminX;
                    int h ← blobs.getMaxY - blobs.getMinY;
                    outfile<<",BB"<<compteur<<compteurblb<<"-"<<(int)blobs.getMinX<<"-" <<(int)
                     blobs.getMinY<<"-"<<w<<"-"<<h;
                    img.getRectangle(blobs.GetCordinate,w,h);
                    img.getCircle(blobs.getCentroid);
                    compteurblb++;
    end for
    outfile ← endl;
    Mat bgimg;
    bg_model.getBackgroundImage(bgimg);
    imshow("image", img);
    imshow("foreground mask", fgmask);
    imshow("foreground image", fgimg);


    if(!bgimg.empty()) then
                        imshow("mean background image", bgimg );
    end if
    nframes++;
    char k ← (char)waitKey(1);
    if ( k = 27 ) break;
    if ( k = ' ' ) then
                update_bg_model ← !update_bg_model;
```

```
        if (update_bg_model) then
            printf("Background update is on\n");
        else
            printf("Background update is off\n");
        end if
    end if
end while.
```

Pour l'analyse vidéo, on a utilisé la librairie OpenCV [79] pour extraire les données de bas niveau des bounding box correspondant à chaque blob comme par exemple : Top_Left_Point_X, Top_Left_Point_Y, Height and Width. Toutes ces données sont alors enregistrées

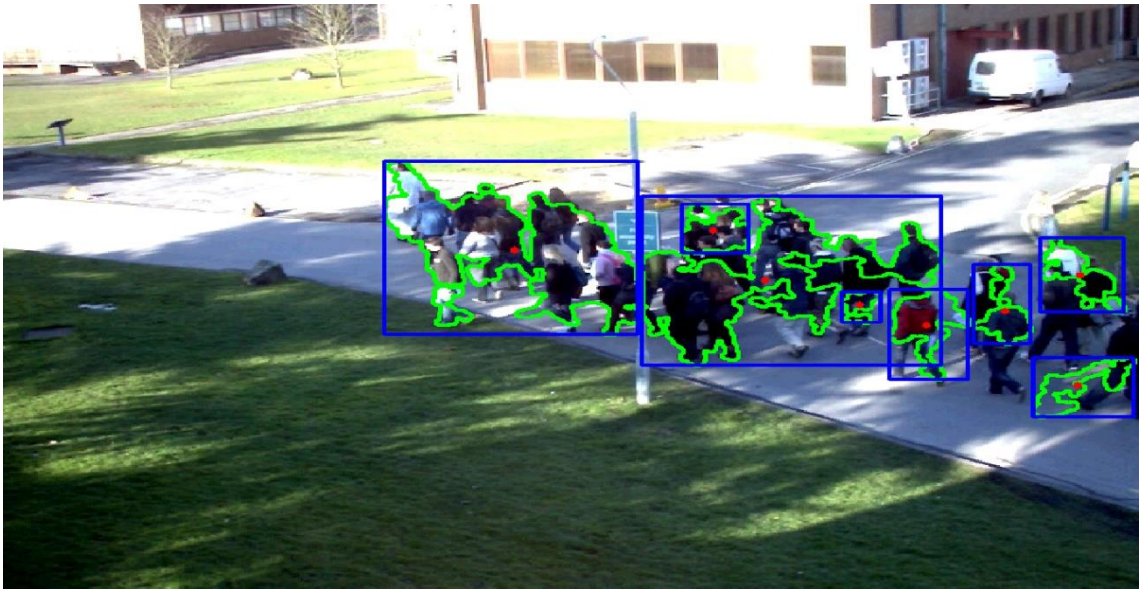


Figure 19. Illustration d'une image d'extraction d'informations de bas niveau.

La Figure 19 illustre un exemple de données extraites lors d'une analyse de bas niveau de notre système d'indexation OVIS utilisant la méthode du background soustraction. Dans la Figure 19, les blobs sont représentés sous forme de pixels délimités par un trait de couleur verte. En général, un blob est représenté par une région de l'image qui est soit plus claire, soit plus sombre que son entourage. Cela veut dire que les pixels de cette région doivent être plus clairs ou plus sombres que leur voisinage. De leurs côtés, les différents bounding box sont représentés sous forme de carrés bleus qui englobent les différents blobs détectés. Notre système d'indexation OVIS utilisent alors les données concernant ces différents bounding box comme informations de base (en entrée du système), afin d'extraire l'évènement correspondant.

4.3.2. Les règles SWRL

Pour tester l'efficacité de notre approche, on a utilisé différents évènements des vidéos des benchemarks PETS 2012 avec plus de 300 règles SWRL, tels que :

- Les évènements Group Running et Walking : dans chaque image, l'ampleur du mouvement identifie la différence entre ces deux évènements. Par exemple, une grande ampleur signifie un mouvement de Group running, alors qu'une petite ampleur signifie un mouvement de Group Walking. La détection est faite par la définition d'un seuil d'expérimentation où on utilise un classifieur avec quelques options telle que la vitesse moyenne du mouvement. Dans notre cas, on a utilisé un seuil d'expérimentation.
- Les évènements Group Formation et Splitting : La position, l'orientation et la vitesse de chaque groupe sont les facteurs importants pour la détermination de l'évènement.
- L'évènement Group Local Dispersion : la position et l'évolution de la taille du groupe à travers les images successives représentent les facteurs clés pour la détermination de l'évènement

Dans notre expérimentation, on a utilisé le plugin des règles SWRL de Protégé 2000 [80] pour l'écriture des règles avec un langage SWRL, et le raisonneur Pellet [81] pour inférer tous les évènements des vidéos. Toutes ses règles SWRL sont classées en trois classes :

a. Règles SWRL de distance

Elles consistent à générer les major bounding box dans chaque image de la séquence vidéo. Ces règles vérifient la distance entre les bounding boxes détecté dans l'image courante. Les Bounding boxes voisins sont groupés en un seul bounding majeur box.

La Figure 20 montre un exemple de construction d'une règle SWRL de distance, avec deux bounding boxes détecté avec le module d'extraction des blobs. Cette règle SWRL vérifie si un groupement de ces bounding boxes peut être réduit à un seul Bounding Box majeur. Le raisonneur Pellet prend la décision d'inférer ou pas la partie droite de la règle en vérifiant la partie gauche de la règle.

Figure 20. Illustration d'une règle SWRL pour le groupement de deux bounding boxes en un seul majeur.

b. Règles SWRL de suivi :

Ces règles consistent à générer les différents groupes de personne (Group_Of_Person) et utilisent le résultat généré par les règles SWRL de distance pour détecter la position de départ et celle d'arrivée de chaque groupe de personnes, ainsi que d'autres paramètres dans la vidéo séquences. La Figure 21 illustre un exemple de la construction d'une règle de traçage SWRL, en décrivant le traçage du groupe nommé GPY entre deux images successives (FZ et FZ+1).

Figure 21. Illustration d'une règle SWRL pour tester si le bounding box majeur détecté dans l'image FZ+1 représente le même groupe GPY détecté dans l'image FZ.

c. Règles SWRL d'évènement :

La troisième catégorie de règles sert à détecter l'évènement approprié en analysant le comportement du groupe identifié dans la catégorie précédente.

Comme un exemple de construction d'une règle d'évènement SWRL, la Figure 22 montre si le groupe nommé GPZ se divise ou non en deux sous-groupes (GPX et GPY) entre deux frames d'images successive (FZ et FZ+1).

Figure 22. Illustration d'une règle SWRL pour le test d'un évènement de division du groupe GPZ.

4.4. Résultats et discussions

Pour tester et valider notre système d'indexation et de recherche des vidéos OVIS, on a expérimenté cinq cas d'études d'identification d'évènements en utilisant des vidéos des benchmarks PETS 2012 et TRECVID 2016. Pour cela, nous avons développé notre application sous JAVA sur une machine avec un processeur Intel Core I7 CPU et 8 GB de RAM, sous Windows 8. Dans notre expérimentation, nous avons considéré trois types d'évaluation afin de tester la performance de notre système d'indexation OVIS.

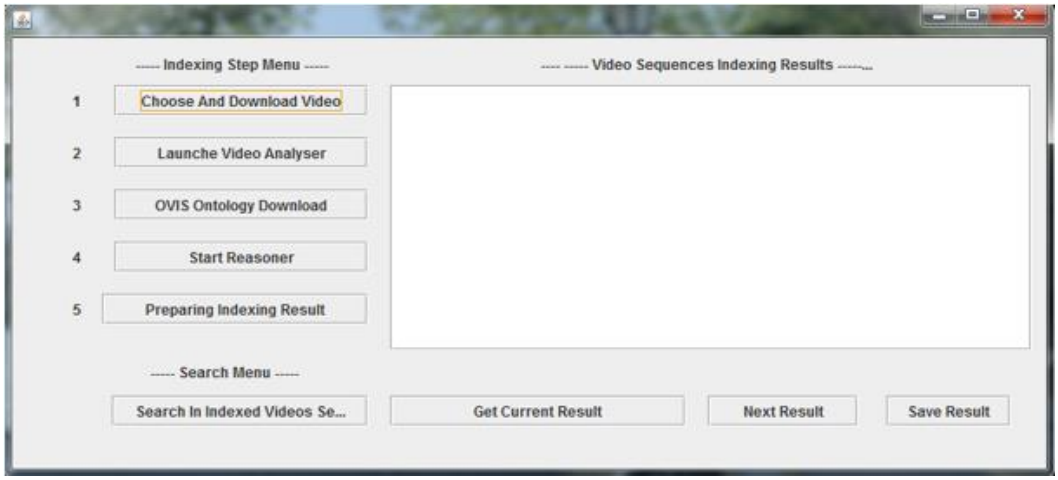


Figure 23. Interface graphique principale de notre système OVIS.

La figure 23 présente l'interface graphique de notre système OVIS, ou on peut sélectionner la vidéo à indexer et lancer l'analyseur de bas niveau pour attribuer donc un index a la vidéo (voir détails d'un cas pratique d'exécution du système en annexe B).

4.4.1. Evaluation basé sur les évènements.

Le premier type d'évaluation est basé sur le nombre d'évènements retourné par notre système OVIS ; cette première évaluation est testée selon plusieurs métriques, telles que : la précision, le rappel, le F-mesure, FP (Faux Positif), FN (Faux Négatif), VP (Vrai Positif) and VN (Vrai Négatif). On considère les métriques comme suit :

- Précision = Nombre de vidéos détectées qui contiennent l'action / nombre de vidéos indexées avec l'action.
- Rappel = Nombre de vidéo détectées qui contiennent l'action / Nombre de tous les vidéos qui contiennent l'action.
- F-mesure = 2 * ((Précision * Rappel) / (Précision + Rappel)).
- VP = Nombre de vidéos indexées avec l'action et qui la contiennent réellement.

- VN = Nombre de vidéos non indexées avec l'action qui ne la contiennent pas réellement.
- FP = Nombre de vidéos indexées avec l'action qui ne la contiennent pas réellement.
- FN = Nombre de vidéos non indexées avec l'action qui la contiennent réellement.

Afin d'évaluer notre système OVIS, on a expérimenter 16 vidéos de 4240 images, du benchmark PETS 2012 en tenant compte de toutes les métriques, et 11h d'une vidéo du benchmark TRECVID 2016 en utilisant seulement les trois premières métriques.

a. Le Benchmark PETS 2012

Résultats / Evènements	Nombre de toutes les vidéos qui représentent réellement l'évènement dans la base de données (vérité terrain)	Nombre de vidéos retournées	Nombre de vidéos retourné qui représentent réellement l'évènement
Walking	13	11	11
Running	08	09	05
Splitting	04	06	03
Formation	04	06	03
Local dispersion	04	04	04

Tableau 7. Résultat d'indexation des différents évènements (cas du PETS 2012).

Le Tableau 7 montre le résultat d'indexation pour chaque évènement. Dans la 2eme colonne du tableau, on considère 13 vidéos de l'évènement Walking, 8 vidéos de l'évènement Running, 4 vidéos de l'évènement Splitting, 4 vidéos de l'évènement Formation et 4 vidéos de l'évènement local dispersion comme vérité terrain. La 3eme colonne du tableau est consacrée au nombre de vidéos indexés par notre système OVIS pour chaque évènement ; la 4eme colonne du tableau montre le nombre de vidéos indexés contenant l'évènement parmi ceux indexés par notre système OVIS.

Mesures / Evènement	Précision	Rappel	F-measure	FP	FN	VP	VN
Walking	100%	84%	91%	00	02	11	03
Running	55%	62%	58%	04	03	05	04
Splitting	50%	75%	60%	03	01	03	09
Formation	50%	75%	60%	03	01	03	09
Local dispersion	100%	100%	100%	00	00	04	12

Tableau 8. Les résultats obtenus pour chaque métrique.

➢ Discussion 1

Le tableau 8 résume les statistiques de données obtenues en utilisant le benchmark PETS 2012. On voit bien que les évènements Walking et Local Dispersion donnent de bons résultats et atteignent un taux de 100 % en termes de précision. Cependant, ces résultats veulent dire que le nombre de vidéos détectés par notre système OVIS et qui contiennent les évènements cités auparavant est égal au nombre de vidéo indexés avec ces mêmes évènements. D'autre part, les

évènements Running, Splitting et Formation dépassent le taux de 50% en termes de précision. Alors, on peut conclure que le nombre de vidéos détectées par notre système OVIS qui contiennent ces évènements est égal au minimum à la moitié des vidéos indexé par ces mêmes évènements. Comme c'est illustré dans le tableau 8, l'évènement Local Dispersion présente aussi un excellent résultat de 100% en termes de rappel. Ceci dit, le système OVIS n'a raté aucun évènement de Local Dispersion. De plus, le rappel des évènements Walking, Running, Splitting et Formation donnent de bons résultats qui dépassent les 62%. Suivant les résultats de précision et rappel, la métrique F-mesure exprime la relation Précision/Rappel. Cependant, la métrique F-mesure fournit un excellent résultat aux évènements Walking et Local Dispersion et de bons résultats dans le reste des évènements. Les métriques (FP, FN, VP et VN) donnent la pertinence du processus d'indexation généré par le système OVIS. Ainsi, cette pertinence est considérée comme excellente lorsque le nombre de FP et FN est très bas et que le nombre de VP et VN est très haut. Comme c'est illustré dans le tableau 8, ces métriques donnent de bons résultats en général. Par exemple, l'évènement Walking génère de bons résultats avec un 3/3 pour (VN), 11/13 pour (VP), 0/3 pour (FP) et 2/13 pour (FN). De ce fait, cela veut dire que notre système OVIS qui utilise les règles SWRL avec les différentes catégories (règles de distance, règles de traçage, et règles d'évènement) sans l'utilisation des méthodes classiques (SVM, KNN, etc…), ouvre de nouvelles perspectives dans le processus d'indexation et de recherche des vidéos, et cela grâce à l'utilisation de notre ontologie de vidéosurveillance.

- **Comparaison de notre système OVIS avec d'autres approches**

Dans le but d'évaluer notre system OVIS qui est basé sur des règles SWRL, on a comparé notre système avec deux autres approches [45, 46] qui utilisent les même séquences vidéo du benchmark PETS 2012 pour la détection des évènements de vidéo, comme par exemple : Walking, Running, Splitting, Formation et Local Dispersion.

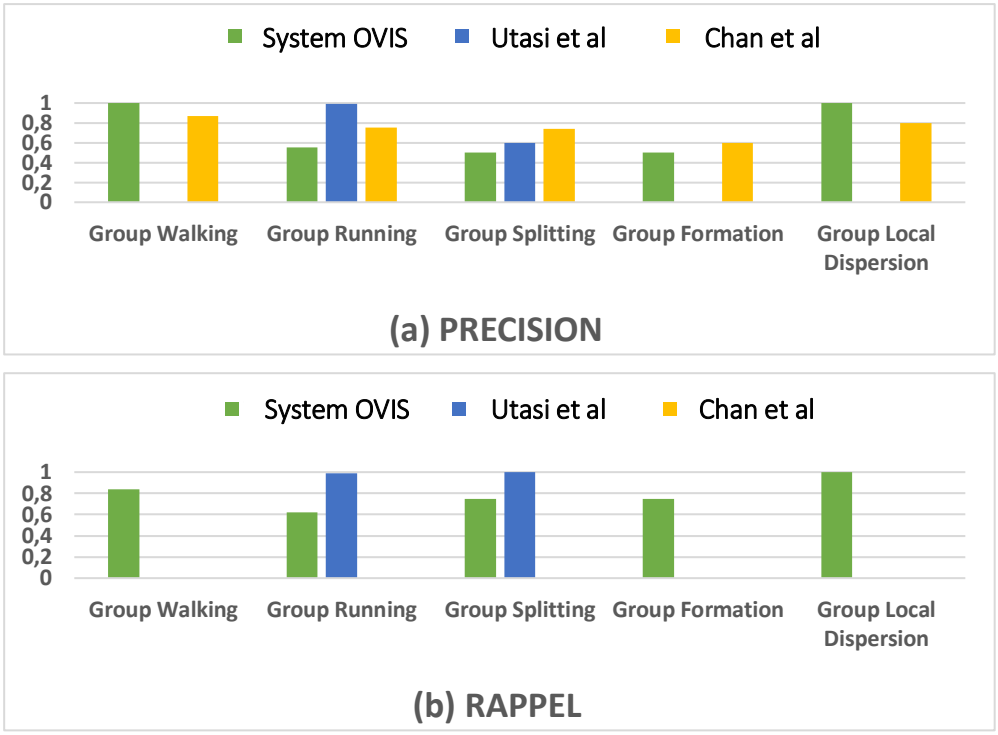


Figure 24. Comparaison de différentes approches de détection d'évènements sous forme graphique, (a) Précision, (b) Rappel (PETS 2012).

Evènement	Métriques	Le system OVIS	Utasi et al. [45]	Chan et al. [46]
Group Walking	Precision	1	ND	0,87
	Recall	0,84	ND	NC
Group Running	Precision	0,55	0,99	0,75
	Recall	0,62	0,99	NC
Group Splitting	Precision	0,5	0,6	0,74
	Recall	0,75	1	NC
Group Formation	Precision	0,5	ND	0,6
	Recall	0,75	ND	NC
Group Local Dispersion	Precision	1	ND	0,8
	Recall	1	ND	NC

Tableau 9. Comparaison de différentes approches de détection d'évènements (PETS 2012), NC (Non Communiqués), ND (Non Détectés).

Le Tableau 9 compare les résultats obtenus avec le système OVIS et ceux reportés dans les travaux [45, 46] en utilisant le benchmark PETS 2012. La Figure 24 au-dessus, illustre la relation Précision/Rappel de notre approche comparée au deux travaux cités auparavant.

Notre approche basée sur les règles SWRL détecte tous les évènements, tandis que la méthode [45] détecte seulement deux évènements (Running et Splitting) et ne permet pas d'identifier les évènements manquants bien qu'ils disent que leur approche est capable de détecter les autres évènements sans donner les moindres détails sur la façon avec laquelle ils procédent. La méthode de Chan et al. [46] ne peut pas traiter deux évènements en même temps

alors que notre approche OVIS offre la possibilité d'exécuter et de détecter deux évènements ou plus au même moment, comme c'est illustré dans la Figure 26 en dessous. De plus, l'approche de Chan et al. [46] ne fournit pas de données très importantes comme c'est mentionné par NC dans le tableau 9. Même si la précision de [46] est meilleure que la nôtre dans quelque cas d'études, les auteurs ne fournissent cependant pas leur métrique Rappel.

Ces observations nous mènent à conclure que ces points négatifs cités plus haut (la non détection de tous les évènements, la gestion d'un seul évènement à la fois et aussi la non présentation du résultat important tel que le Rappel), prouvent que notre approche basée sur les règles SWRL est très efficace. Cependant, la possibilité d'ajouter de nouveaux évènements est facile grâce à la création de nouvelles règles d'inférence.

b. Le Benchmark TRECVID 2016

Afin d'évaluer notre system OVIS selon des contextes différents, on a fait une comparaison de notre approche avec celles de [48, 49] qui ont contribué à SED (Surveillance Event Detection) de TRECVID 2016. Ce dernier regroupe sept évènements, à savoir : PersonRuns, CellToEar, ObjectPut, PeopleMeet, People-SplitUp, Embrace, et Pointing. Dans notre cas d'étude, les règles SWRL de notre ontologie de vidéosurveillance OVIS traite trois évènements parmi les sept, à savoir : PersonRuns (Running), PeopleMeet (Formation), PeopleSplitUp (Splitting), avec lesquels on les a comparés avec les résultats obtenus par OVIS.

Results / Events	Number-of- event (ground truth)	Number-of-all-event-returned-by-OVIS	Number-of-correct-event-returned- by-OVIS
People Meet	323	411	303
People Split Up	176	203	173
Person Runs	63	97	58

Tableau 10. Le résultat d'indexation des différents évènements (TRECVID 2016).

Le tableau 10 montre les résultats d'indexation des évènements comme par exemple : PeopleMeet, PeopleSplitUp et PersonRuns, où le system OVIS utilise les règles d'inférence SWRL pour fournir les évènements suivants : Formation, Splitting et Running. Dans nos expérimentations, on a considéré 323 évènements pour PeopleMeet, 176 pour PeopleSplitUp et 63 pour PersonRuns comme vérité terrain (Ground-Truth) (voir première colonne du tableau 10). Chaque évènement a été détecté donc avec son image de début et de fin, ceci permet alors de confirmer une bonne détection d'évènement ou une fausse alarme.

Event	Metrics	The OVIS System	Markatopoulou et al. [48]	Zhao et al. [49]
People Meet	Precision	0,74	0,02	0,34
	Recall	0,94	0,92	0,18
	F-measure	0,83	0,04	0,23
People Split Up	Precision	0,86	0,01	0,32
	Recall	0,98	0,98	0,2
	F-measure	0,92	0,02	0,25
Person Runs	Precision	0,6	0,01	0,67
	Recall	0,92	0,97	0,35
	F-measure	0,73	0,02	0,46

Tableau 11. Comparaison des différentes approches de détection d'évènements en utilisant le benchmark TRECVID 2016.

Le tableau 11 montre la première comparaison des résultats obtenus par notre système OVIS et ceux obtenus dans [48, 49] en utilisant l'évènement SED du benchmark TRECVID 2016. La Figure 25 quant à elle montre la deuxième comparaison de notre système OVIS toujours avec les approches [48, 49] selon les métriques : Précision et Rappel.

Figure 25. Comparaison des différentes approches de détection d'évènements sous forme graphique en utilisant le benchmark TRECVID 2016, (a) Précision, (b) Rappel.

➢ **Discussion des résultats**

- Basée sur les règles SWRL, Notre approche génère un meilleur rapport Précision/Rappel (F-measure) comparativement aux approches [48, 49], ce qui veut dire que notre system OVIS détecte la majorité des évènements sans pour autant générer un nombre important de fausses alarmes.

- L'approche de Markatopoulou et al. [48] détecte la majorité des évènements correctes tout comme notre system OVIS, mais simultanément avec un grand nombre de fausses alarmes. On peut dire qu'une telle approche ne prend pas en charge un grand nombre de détections d'évènements correctes, mais la précision de détection est beaucoup moins importante que celle générée par le system OVIS. En outre, l'approche de Zhao et al. [49] présente des résultats acceptables en termes de précision mais malheureusement elle néglige un nombre important de détections d'évènements corrects comparativement à notre system OVIS.
- 3- Un point fort de notre système OVIS comparativement aux approches de [48, 49], est qu'il permet d'ajouter facilement de nouveaux évènements comme par exemple : CellToEar, ObjectPut, Embrace, Pointing, …, et ceci en créant leurs propres règles SWRL.
- En conclusion, nous pouvons dire que toutes les limites et lacunes ayant été détectées dans les approches [48, 49] comme la négligence d'un nombre important de détection d'évènements correctes et de fausses alarmes, prouvent que notre système OVIS à base des règles SWRL est plus efficace et très performant.

4.4.2. Evaluation basée sur la temporisation en images

Le deuxième type d'évaluation est basé sur la temporisation en images, en utilisant un seuil (exprimé en nombre d'images, 10 images dans notre cas), pour évaluer la performance de notre system OVIS. Pour cela, trois cas sont pris en compte :

- **Cas 1 : Trop tôt :** notre system détecte l'évènement avant son commencement réel (vérité terrain) avec 10 images.
- **Cas 2 : Dans le temps :** notre system détecte l'évènement exactement lors de son commencement.
- **Cas 3 : Trop tard :** notre system détecte l'évènement après son commencement réel (vérité terrain) avec 10 images.

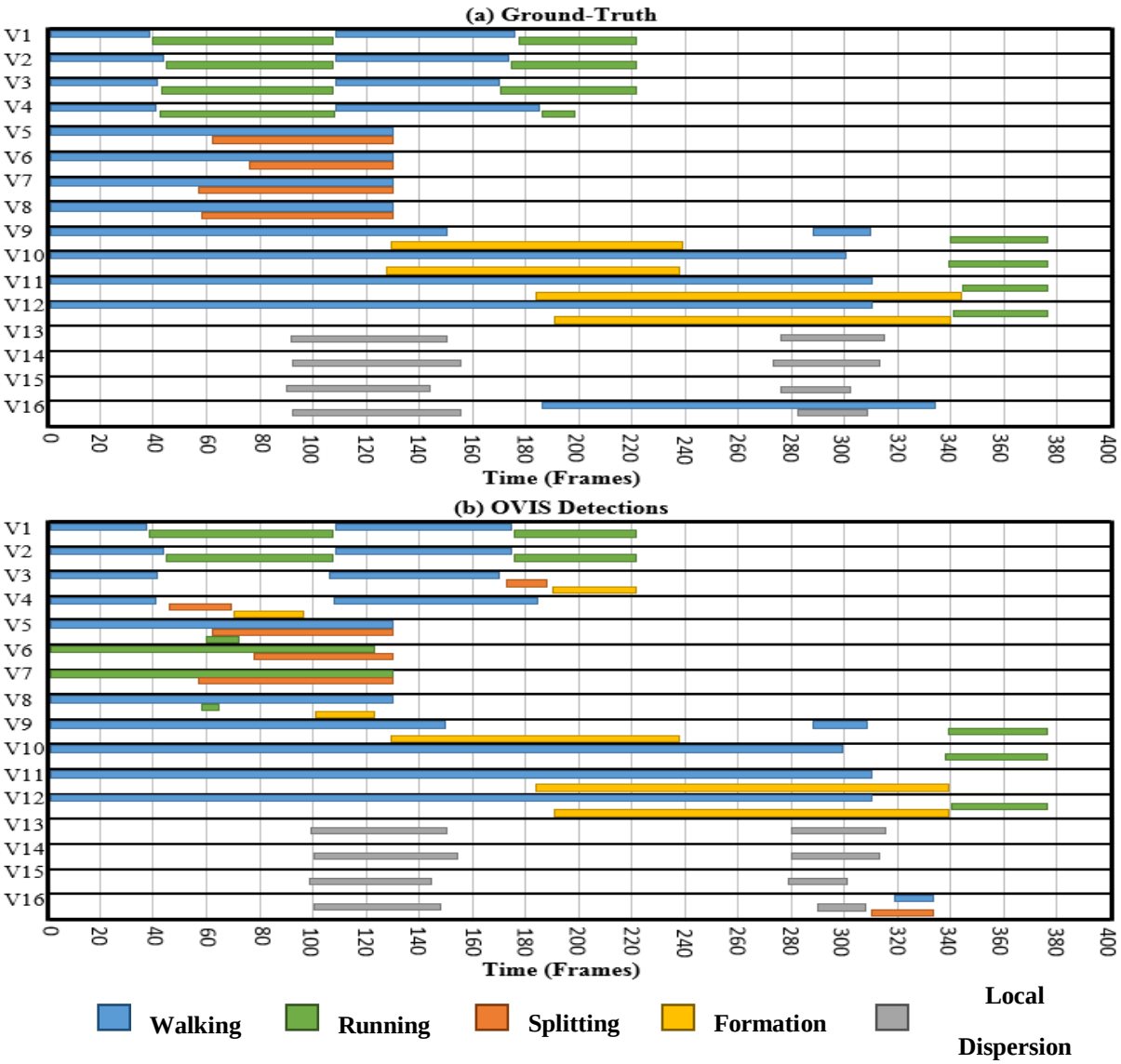


Figure 26. Temporisation en images d'évènements des 16 vidéos (vérité terrain et résultats obtenus par le system OVIS).

La Figure 26 représente la temporisation en image des évènements des 16 vidéos (vérité terrain et résultats retournés par le système OVIS). L'annotation pour la vérité terrain est faite manuellement en regardant la vidéo en entier, et représente dans chaque cas l'évènement qui se produit avec son début et sa fin.

➢ **Discussion des résultats**

La Figure 26 illustre les résultats de la temporisation en images des évènements obtenus par notre système OVIS en comparaison avec la vérité terrain. Les avantages de notre système OVIS sont comme suit :

- La détection d'au moins un évènement correct dans une séquence vidéo.

- Tous les évènements détectés correctement, ont été ajustés dans le temps sans dépasser le seuil de 10 images.

Ces avantages prouvent que notre système OVIS basé sur les règles SWRL fonctionne correctement et détecte l'évènement correct au moment voulu comparativement aux approches [45, 46, 48, 49]. Néanmoins, l'inconvénient majeur de notre système OVIS, est la confusion occasionnelle d'évènements. Ceci est dû généralement à deux raisons :

- L'émergence d'autres objets (non pertinents) détectés dans la séquence vidéo par le module d'analyse de bas niveau. Ainsi, leur comportement induit notre système en erreur pour détecter de faux évènements. Par exemple, dans le cas où le module d'analyse vidéo détecte un mouvement d'ombre d'arbre et le considère comme Bounding Box ce qui représente une fausse détection et mène-le système OVIS à générer de faux évènements.
- La détection incorrecte d'évènements de dispersion totale et de formation à cause de la vitesse de marche ou de la course des objets détectés. Par exemple, lorsque dans le même groupe de personne, il s'agit de différentes vitesses de marche de personnes, notre système OVIS peut détecter de faux évènements de dispersion totale et de formation.

Nous pouvons dire que tous les inconvénients de notre système OVIS peuvent être résolus dans nos travaux futurs, en créant de nouvelles règles SWRL.

4.4.3. Evaluation basée sur l'algèbre d'intervalles d'Allen

Le troisième type d'évaluation est basé sur l'algèbre d'intervalles d'Allen [82]. Ses différents intervalles schématisent une représentation temporelle concrète associée à des règles de déduction (*avant, après, pendant*, …). Généralement, ils représentent des relations et des scénarios entre les différents évènements et assurent leurs cohérences afin de modéliser plusieurs situations possibles comme " X *vient avant* Y", "X *chevauche* Y", "X *rencontre* Y", etc... (Où X et Y représentent deux évènements différents détectés dans la même séquence vidéo). En outre, ces intervalles d'Allen possèdent un temps de début et un temps de fin d'évènements et sont représentés en 5 zones comme l'illustre la figure 27 (Z1 : avant l'évènement, Z2 : début de l'évènement, Z3 : pendant l'évènement, Z4 : fin de l'événement, Z5 : après l'évènement).

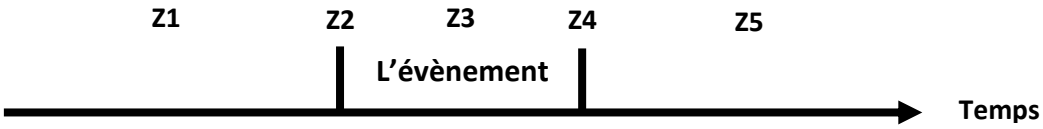


Figure 27. Illustration temporelle des 5 zones représentant un événement.

Pour traduire ces intervalles d'Allen dans notre système OVIS, un autre type de règles SWRL a été créé afin d'inférer différentes relations entre les évènements détectés dans une même séquence vidéo. Les résultats de ces règles SWRL sont sauvegardés comme une Data_Property de type Booléen relative aux individus issus du concept Video_Sequences (comme exemple la relation : Took_Place_Before_R). Cependant, toutes les relations exprimées comme Data_Property inférées avec ce nouveau type de règles SWRL prennent les valeurs vraies ou fausses.

Pour inférer les différentes relations entre les évènements détectés dans une même séquence vidéo, les images de début et de fin de l'évènement représentent alors des aspects clé. De plus, la détection de 5 images entre deux évènements successifs différents est considérée dans notre cas comme les relations "Met_R / Met_By_R", et la détection de plus de 5 images comme les relations : "Took_Place_Before_R / Took_Place_After_R". Cependant, lorsqu'une séquence vidéo contient seulement un évènement qui est détecté plusieurs fois (cas des vidéos 13, 14, et 15 du Tableau 12), aucune relation n'est alors inférée.

En résumé, il existe 7 relations possibles et leurs inverses :

- X took place before Y : Lorsque X finit avant le début de Y.
- Y took place after X : Lorsque Y débute après que X finisse.
- X met Y : Lorsque X finit au même moment que le début de Y.
- Y met by X : Lorsque Y débute au même moment que X finisse.
- X overlapped Y : Lorsque X débute avant Y et que Y finit après X.
- Y overlapped by X : Lorsque Y débute après X et que X finit avant Y.
- X started Y : Lorsque X débute au même moment que Y et finit avant Y.
- Y started by X : Lorsque Y débute au même moment que X et finit après X.
- X took place during Y : Lorsque X débute après Y et finit avant Y.
- Y contained X : Lorsque Y débute avant X et finit après X.
- X finished Y : Lorsque X débute après Y et finit au même moment que Y.
- Y finished by X : Lorsque Y débute avant X et finit au même moment que X.
- X equalled to Y : Lorsque X débute et finit au même moment que Y.

Sequences Vidéos	Les relations d'Allen détecter par OVIS
Vidéo 1	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Vidéo 2	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Vidéo 3	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Vidéo 4	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R
Vidéo 5	Finished_R / Finished_By_R / Overlapped_R / Overlapped_By_R / Took_Place_During_R / Contained_R
Vidéo 6	Overlapped_R / Overlapped_By_R
Vidéo 7	Finished_R / Finished_By_R
Vidéo 8	Took_Place_During_R / Contained_R / Took_Place_Before_R / Took_Place_After_R
Vidéo 9	Overlapped_R / Overlapped_By_R / Took_Place_Before_R / Took_Place_After_R
Vidéo 10	Took_Place_Before_R / Took_Place_After_R
Vidéo 11	Overlapped_R / Overlapped_By_R
Vidéo 12	Overlapped_R / Overlapped_By_R / Met_R / Met_By_R
Vidéos 13 /14/ 15	
Vidéo 16	Took_Place_Before_R / Took_Place_After_R / Met_R / Met_By_R

Tableau 12. Illustration des résultats d'inférence représentant les relations d'Allen générés par notre system OVIS.

- **Discussion des résultats**

Le Tableau 12 illustre toutes les relations d'Allen générées par notre système OVIS. Ainsi, le premier avantage de notre système est la détection correcte de toutes les relations d'Allen au niveau des 16 vidéos. Par exemple, la vidéo 12 "V12" de la Figure 26 (partie **OVIS Detections**) comporte trois types d'évènements : *marcher, formation et courir*.

On peut alors inférer les différentes relations d'Allen à partir d'évènements cités précédemment comme suit :

- L'évènement marcher débute avant l'évènement formation et l'évènement formation finit après l'évènement marcher, ce qui traduit les relations d'Allen suivants : (X overlapped Y) et (Y overlapped by X),
- L'évènement formation finit au moment du début de l'évènement courir, ce qui traduit les relations d'Allen suivants : (X meet Y) et (Y met by X).

Cependant, toutes ces relations d'Allen sont inférées par notre système OVIS comme c'est illustré dans le tableau 12 (cas de la vidéo 12).

Le deuxième avantage est que notre système OVIS ne détecte aucune relation dans le cas d'une séquence vidéo ne comportant qu'un seul type d'évènement, comme c'est le cas de l'évènement **Local-Dispersion** des vidéos : 13, 14 et 15 illustrées dans la Figure 26, et aucune relation n'a été inférée par notre système OVIS comme l'illustre le Tableau 12.

Tous ces avantages montrent que notre approche utilisant les règles SWRL tient compte de toutes les relations d'Allen contrairement aux approches [45, 46, 48, 49].

4.5. Conclusion

Au cours des dernières années, beaucoup d'utilisateur se sont intéressé à la recherche d'informations au niveau de plusieurs secteurs d'activités académiques et industrielles. Concernant le domaine de la vidéo surveillance, l'approche ontologique a prouvé son efficacité du fait qu'elle peut être utilisée dans le processus d'indexation pour la détection d'évènements. L'ontologie permet de modéliser un ensemble de connaissances et par conséquent, elle est considérée comme un outil de représentation de description de bas niveau des vidéos.

Dans ce chapitre, nous avons présenté notre système d'indexation et de recherche des vidéos appelé OVIS (Ontology Vidéo-Surveillance Indexing System) [6]. Notre système OVIS utilise un module de bas niveau pour extraire les blobs dans des séquences vidéo et un moteur d'inférences à bases de règles SWRL pour inférer les évènements détectés. Pour cela, trois types de règles ont été utilisées : les règles de distance, les règles de traçage et les règles d'évènement. Nous avons aussi présenté les résultats des expérimentations obtenus en utilisant des vidéos des benchmarks universels comme PETS 2012 et TRECVID 2016 ainsi que les différentes évaluations : celles basées sur les évènements, celles basées sur la temporisation en images et les autres basées sur l'algèbre d'intervalle d'Allen.

On peut dire que les résultats des expérimentations de notre système d'indexation et de recherche des vidéos OVIS étaient plus que satisfaisants comparativement à ceux obtenus par les approches de l'état de l'art.

Conclusion Générale et Perspectives Futures

De nos jours, les systèmes de vidéosurveillance sont devenus une partie essentielle de notre vie quotidienne, à cause de leur rôle en termes d'assurance en sécurité (i.e. permettant l'étude du comportement humain au sein de la population). Par ailleurs, les avancées technologiques dans le domaine numérique ont aussi engendré une quantité très importante de documents vidéo, et cela grâce à la diversité d'outils de captures tel que (les téléphones portables, les caméras, les appareils photos, les tablettes, etc.).

Par ailleurs, l'indexation et la recherche des vidéos dans une vidéosurveillance est d'un intérêt majeur chez la communauté des chercheurs dans le domaine de la vision par ordinateur. En effet, on s'intéresse davantage à l'analyse du comportement humain depuis des vidéos capturées dans un endroit publique comme : les aéroports, les gares ferroviaires, les parcs de voitures, les supermarchés, etc.

Plusieurs travaux de recherche existent ou sont en cours de développement. Ils s'intéressent au domaine de la vidéosurveillance, mais il n'existe pas toujours un système d'indexation et de recherche efficace pour traiter un grand volume de vidéos.

Au niveau de plusieurs secteurs d'activités, la recherche d'information vidéo est une tâche nécessaire qui répond aux besoins d'un grand nombre d'utilisateurs du multimédia. Cependant, la majorité des travaux de recherche sont basés sur les aspects technologiques comme l'analyse et l'extraction automatique de l'information audiovisuelle et ne traitent pas donc les vrais besoins des utilisateurs. Pour cela, les ontologies sont devenues plus que nécessaires pour résoudre le problème de sémantique et ont été introduites dans les nouvelles méthodes et techniques de recherche d'information multimédia afin de satisfaire les vrais besoins de l'utilisateur.

En général, les ontologies sont un outil de l'intelligence artificielle (IA), qui sert à modéliser les connaissances d'un domaine particulier comme celui de la vidéosurveillance dans notre cas. Ainsi, son intégration dans l'indexation et la recherche des vidéos a vu le jour ces dernières années et plusieurs systèmes d'indexation des vidéos de surveillance existent actuellement sur le marché.

Dans notre travail de recherche, nous avons proposé et développé une ontologie formelle (syntaxe et sémantique) de vidéosurveillance et une approche basée sur les règles d'inférence SWRL pour détecter les évènements d'objets multiples ou éventuellement les évènements de foule de gens. Par exemple un groupe de gens qui marchent, ou un groupe de gens qui se dispersent totalement ou localement, etc. Notre ontologie de vidéosurveillance ainsi proposée décrit la liaison entre les quatre grandes catégories de vidéos, à savoir : Video_Sequences, Video_Objects, Video_Events et Video_Actions, qui sont en interaction entre elles.

Elle couvre un nombre important d'objets, d'actions et d'évènement de la vidéosurveillance et elle peut être utilisée à large échelle dans le domaine industriel.

En outre, on a conçu et réalisé un système d'indexation et de recherche des vidéos basé sur l'ontologie proposée, appelé OVIS (Ontology Video Surveillance Indexing and Retrieval system) en utilisant des règles SWRL. Notre système OVIS a été expérimenté avec des vidéos des benchmarks universels PETS 2012 et TRECVID 2016, et il a été évalué avec des évènements d'objets multiples (Ex. Groupe qui marche, Groupe qui court, etc.). Les résultats obtenus ont été très satisfaisants comparativement à ceux obtenus dans les systèmes d'indexation de l'état de l'art.

Lors du processus d'indexation, notre système OVIS utilise des règles SWRL au niveau moyen et haut niveau des vidéos et permet de créer facilement de nouvelles règles pour d'autres besoins d'utilisateurs de la vidéosurveillance. Il peut aussi être appliqué à d'autres domaines tels : que la création des benchmarks, la description de scène, etc...

Notre travail de thèse est loin d'être achevé. Pour cela, on envisage les perspectives futures suivantes :

- Une extension de notre system OVIS en considérant de nouveaux évènements de la vidéosurveillance, en ajoutant de nouvelles règles SWRL.
- Expérimentation de notre système OVIS en utilisant d'autres benchmarks comme TRECVID 2017 et CAVIAR.
- Utiliser le formalisme des réseaux de neurones pour la prédiction automatique d'autres règles SWRL au lieu de les créer manuellement à chaque fois.
- Utiliser la fonction d'entropie normalisée de Shannon pour modéliser l'incertitude associée au module d'extraction d'information de bas niveau d'une vidéo.

Annexe A : Quelques Règles SWRL

Cet Annexe A présente quelques règles SWRL utilisées par notre système OVIS :

Règles SWRL de Distance

Frame(?Fc), BB(?BBa), MBB(?MBBb), BB_Detected_In(?BBa, ?Fc), MBB_Detected_In(?MBBb, ?Fc), Represented_MBB(?BBa, ?MBBb), BB_Top_Left_Point_X(?BBa, ?d), BB_Top_Left_Point_Y(?BBa, ?e), BB_Width(?BBa, ?f), BB_Length(?BBa, ?g) --> MBB_Top_Left_Point_X(?MBBb, ?d), MBB_Top_Left_Point_Y(?MBBb, ?e), BB_Width(?MBBb, ?f), MBB_Length(?MBBb, ?g)

Frame(?Fz), BB(?BBa), BB(?BBb), BB(?BBc), BB(?BBd), BB(?BBe), BB_Detected_In(?BBa, ?Fz), BB_Detected_In(?BBb, ?Fz), BB_Detected_In(?BBc, ?Fz), BB_Detected_In(?BBd, ?Fz), BB_Detected_In(?BBe, ?Fz), BB_Top_Right_Point_X(?BBa, ?p), BB_Top_Right_Point_X(?BBb, ?q), BB_Top_Right_Point_X(?BBc, ?r), BB_Top_Right_Point_X(?BBd, ?s), BB_Top_Right_Point_X(?BBe, ?t), Number_BB_In_Frame(?Fz, 5), swrlb:greaterThan(?p, ?q), swrlb:greaterThan(?q, ?r), swrlb:greaterThan(?r, ?s) , swrlb:greaterThan(?s, ?t) --> BB_ID(?BBa, 1), BB_ID(?BBb, 2), BB_ID(?BBc, 3), BB_ID(?BBd, 4), BB_ID(?BBe, 5)

Frame(?Fc), BB(?BBa), BB(?BBb), BB_Detected_In(?BBa, ?Fc), BB_Detected_In(?BBb, ?Fc), MBB(?MBBa), MBB_Detected_In(?MBBa, ?Fc), BB_ID(?BBa, 1), BB_ID(?BBb, 2), MBB_ID (?MBBa, 1), Number_BB_In_Frame(?Fc, 2), BB_Top_Right_Point_X(?BBa, ?d), BB_Top_Left_Point_X(?BBa, ?e), BB_Top_Right_Point_X(?BBb, ?f), BB_Top_ Left _Point_X(?BBb, ?g), BB_Width(?BBa, ?h), BB_Width(?BBb, ?i), swrlb:add(?j, ?h, ?i), swrlb:add(?k, ?j, 60), swrlb:subtract(?l, ?e, ?f), swrlb:add(?m, ?l, ?j), swrlb:lessThan(?m, ?k), BB_Top_Left_Point_Y(?BBa, ?n), BB_Bottom_Left_Point_Y(?BBa, ?o), BB_Top_Right_Point_Y(?BBb, ?p), BB_Bottom_Right_Point_Y(?BBb, ?q), swrlb:greaterThan(?p , ?o), swrlb:greaterThan(?e, ?g), BB_Lenght(?BBa, ?r), BB_Lenght(?BBb, ?s), swrlb:subtract(?t, ?q, ?n), swrlb:add(?u, ?r, ?s), swrlb:add(?v, ?u, 60), swrlb:lessThan(?t, ?v), swrlb:substract(?w, ?d, ?g) --> MBB_Top_Left_Point_X(?MBBa, ?g), MBB_Top_Left_Point_Y(?MBBa, ?n), MBB_Length(?MBBa, ?t), MBB_Width(?MBBa, ?w)

Frame(?Fc), TBB(?TBBa), MBB(?MBBb), TBB_Detected_In(?TBBa, ?Fc), MBB_Detected_In(?MBBb, ?Fc), Represented_TMBB(?TBBa, ?MBBb), TBB_Top_Left_Point_X(?TBBa, ?d), TBB_Top_Left_Point_Y(?TBBa, ?e), TBB_Width(?TBBa, ?f), TBB_Length(?TBBa, ?g) --> MBB_Top_Left_Point_X(?MBBb, ?d), MBB_Top_Left_Point_Y(?MBBb, ?e), MBB_Width(?MBBb, ?f), MBB_Length(?MBBb, ?g)

Figure 28. Une illustration d'une situation pour regrouper deux Bounding boxes en un seul majeur

Règles SWRL de Suivi

Frame(?Fc), MBB(?MBBd), MBB_Detected_In(?MBBd, ?Fc), MBB_Top_Left_Point_X(?MBBd, ?a), MBB_Top_Left_Point_Y(?MBBd, ?b), MBB_Width(?MBBd, ?c), MBB_Length(?MBBd, ?d), swrlb:divide (?e, ?c, 2), swrlb:divide (?f, ?d, 2), swrlb:add(?g, ?e, ?a), swrlb:add(?h, ?f, ?b) --> MBB_Center_X(?MBBd, ?g), MBB_Center_Y(?MBBd, ?h)

Frame(?Fc), Video_Sequences(?VSa), Frame_Detected_In(?Fc, ?VSa), MBB(?MBBa), TGP(?TGPa), MBB_Detected_In(?MBBa, ?Fc), TGP_Detected_In(?TGPa, ?Fc), Represent_MTGP(?MBBa, ?TGPa), MBB_Center_X(?MBBa, ?g), MBB_Center_Y(?MBBa, ?h), MBB_Width(?MBBa, ?i), MBB_Width(?MBBa, ?j) --> TGP_Center_X(?TGPa, ?g), TGP_Center_Y(?TGPa, ?h), TGP_Width(?TGPa, ?i), TGP_Length(?TGPa, ?j)

Frame(?Fc), Frame(?Fd), Frame(?Fe), Compare_F(?Fc, ?Fd), Number_Frame(?Fc, ?y1), Number_Frame(?Fd, ?y2), Number_Frame(?Fe, ?y3), swrlb:add(?y4, ?y1, 1), swrlb:add(?y5, ?y3, 1), swrlb:equal(?y2, ?y3), swrlb:equal(?y4, ?y5), GP(?GPa), TGP(?TGPa), TGP(?TGPb), MBB(?MBBa), MBB_ID(?MBBa, ?z1), TGP_ID(?TGPb, ?z1), TGP_Detected_In(?TGPa, ?Fc), TGP_Detected_In(?TGPb, ?Fd), Represented_GP(?TGPa, ?GPa), MBB_Detected_In(?MBBa, ?Fd), Check_If_Group(?TGPa, true), MBB_Center_X(?MBBa, ?a), MBB_Center_Y(?MBBa, ?b), TGP_Center_X(?TGPa, ?c), TGP_Center_Y(?TGPa, ?d), swrlb:add(?a1, ?a, 50), swrlb:add(?b1, ?b, 50), swrlb:lessThanOrEqual(?c, ?a1),

swrlb:lessThanOrEqual(?d, ?b1) --> Represent_MTGP(?MBBa, ?TGPb), Compare_F(?Fd, ?Fe), Represented_GP(?TGPb, ?GPa), GPW(?GPa, true)

Frame(?Fc), Number_Frame(?Fc, ?x1), Frame(?Fd), Number_Frame(?Fd, ?x2), swrlb:add(?x3, ?x1, 1), swrlb:equal(?x2, ?x3), GP(?GPa), TGP(?TGPa), GP(?GPb), TGP(?TGPb), GP_Detected_In(?GPa, ?Fc), TGP_Detected_In(?TGPa, ?Fc), GP_Detected_In(?GPb, ?Fd), TGP_Detected_In(?TGPb, ?Fd), Represented_GP(?TGPa, ?GPa), Represented_GP(?TGPb, ?GPb), Check_Relation(?GPa, ?GPb), TGP_Center_X(?TGPa, ?e), TGP_Center_Y(?TGPa, ?f), TGP_Width(?TGPa, ?g), TGP_Length(?TGPa, ?h), TGP_Center_X(?TGPb, ?a), TGP_Center_Y(?TGPb, ?b), TGP_Width(?TGPb, ?c), TGP_Length(?TGPb, ?d), swrlb :lessThan(?d, ?h), swrlb:divide(?i, ?g, 2), swrlb:add(?j, ?e, ?i), swrlb:subtract(?k, ?e, ?i), swrlb:greaterThan(?a, ?k), swrlb:lessThan(?a, ?j), swrlb:divide(?l, ?h, 2), swrlb:add(?m, ?f, ?l), swrlb:subtract(?n, ?f, ?l), swrlb:greaterThan(?b, ?n), swrlb:lessThan(?b, ?m) --> Splited_F_Group(?GPa, ?GPb)

Figure 29. Une illustration d'une situation de vérification si le bounding box majeur détecté dans le frame FZ+1 représente le même groupe GPY détecté dans le frame FZ

Règles SWRL d'Evènement

Video_Sequences(?VSa), GP(?GPa), GP_Detected_In(?GPa, ?VSa), GPW(?GPa, true), Started_F(?GPa, ?x1), Ended_F(?GPa,?x2) --> Walked(?VSa), Started_F_Walked_Event(?VSa, ?x1), Ended_F_ Walked _Event(?VSa, ?x2)

Video_Sequences(?VSa), GP(?GPa), GP(?GPb), GP_Detected_In(?GPa, ?VSa), GP_Detected_In(?GPb, ?VSa), Splited_F_Group(?GPa, ?GPb), Started_F(?GPb, ?x1), Ended_F(?GPb,?x2) --> Split(?VSa), Started_F_Split_Event(?VSa, ?x1), Ended_F_ Split_Event(?VSa, ?x2)

Video_Sequences(?VSa), GP(?GPa), GP(?GPb), GP_Detected_In(?GPa, ?VSa), GP_Detected_In(?GPb, ?VSa), Formed_Group(?GPa, ?GPb), Started_F(?GPb, ?x1), Ended_F(?GPb,?x2) --> Formed(?VSa), Started_F_Formed_Event(?VSa, ?x1), Ended_F_Formed_Event(?VSa, ?x2)

Video_Sequences(?VSa), GP(?GPa), GP_Detected_In(?GPa, ?VSa), GPLD(?GPa, true), Started_F(?GPa, ?x1), Ended_F(?GPa,?x2) --> Local_Dispersed(?VSa), Started_F_Local_Dispersed_Event(?VSa, ?x1), Ended_F_ Local_Dispersed _Event(?VSa, ?x2)

Figure 30. Une illustration d'une situation pour vérifier si le groupe GPZ se divise ou pas

Règles SWRL des Relations d'Allen

Video_Sequences(?VSa), Formed(?VSa), Started_F_Formed_Event(?VSa, ?x1), Ended_F_Formed_Event(?VSa, ?x2), Split(?VSa), Started_F_Split_Event(?VSa, ?x3), Ended_F_Split_Event(?VSa, ?x4), swrlb:add(?x5, ?x2, ?5),swrlb:greaterThan(?x3, ?x5) --> Took_Place_Before_R(?VSa, true), Took_Place_After_R(?VSa, true)

Video_Sequences(?VSa), Local_Dispersed(?VSa), Started_F_Local_Dispersed_Event(?VSa, ?x1), Ended_F_Local_Dispersed_Event(?VSa, ?x2), Formed(?VSa), Started_F_Formed_Event(?VSa, ?x3), Ended_F_Formed_Event(?VSa, ?x4), swrlb:add(?x5, ?x2, ?5), swrlb:lessThanOrEqual(?x3, ?x5), swrlb:subtract(?x6, ?x2, ?5),

swrlb:greaterThanOrEqual(?x3, ?x6), swrlb:greaterThan(?x4, ?x2) --> Met_R(?VSa, true), Met_By_R(?VSa, true)

Video_Sequences(?VSa), Walked(?VSa), Started_F_Walked_Event(?VSa, ?x1), Ended_F_Walked_Event(?VSa, ?x2), Ran(?VSa), Started_F_Ran_Event(?VSa, ?x3), Ended_F_Ran_Event(?VSa, ?x4), swrlb:subtract(?x6, ?x2, ?5), swrlb:greaterThan(?x3, ?x1), swrlb:greaterThan(?x4, ?x2), swrlb:lessThan(?x3, ?x6) --> Overlapped_R(?VSa, true), Overlapped_By_R(?VSa, true)

Video_Sequences(?VSa), Walked(?VSa), Started_F_Walked_Event(?VSa, ?x1), Ended_F_Walked_Event(?VSa, ?x2), Ran(?VSa), Started_F_Ran_Event(?VSa, ?x3), Ended_F_Ran_Event(?VSa, ?x4), swrlb:equal(?x1, ?x3), swrlb:greaterThan(?x2, ?x4) --> Started_R(?VSa, true), Started_By_R(?VSa, true)

Annexe B : Cas pratique de mise en œuvre de notre système OVIS

Cet Annexe B présente un cas pratique d'un processus d'indexation et de recherche des vidéos en utilisant notre système d'indexation OVIS. En premier lieu, les figures 31 et 32 illustrent le démarrage du système d'indexation et consiste à charger les vidéos à indexer.

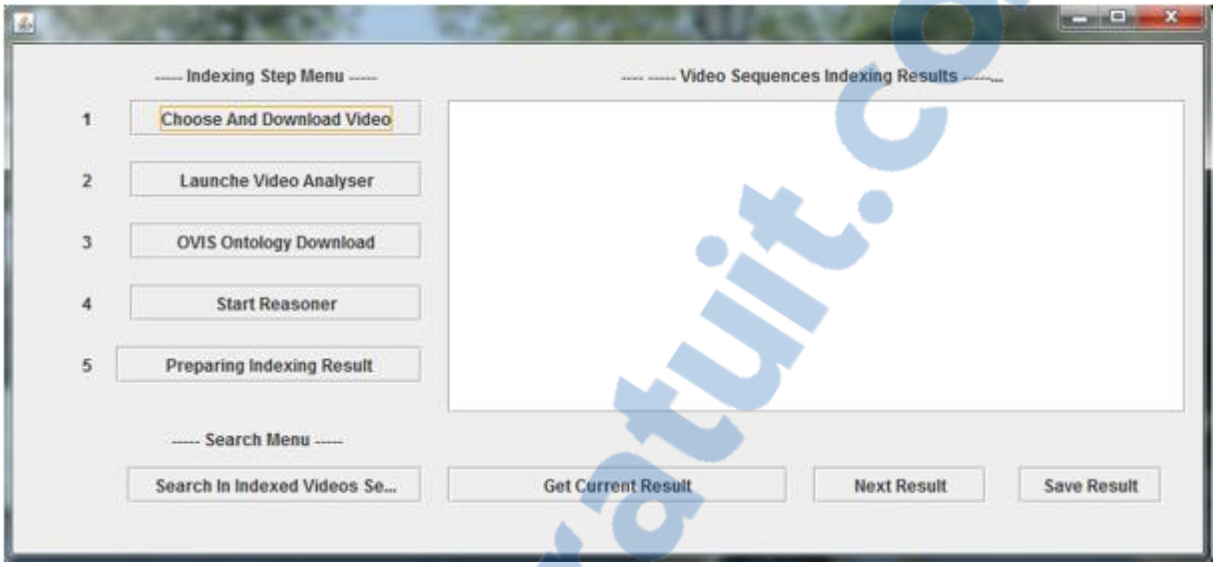


Figure 31. Sélection du bouton Choose et chargement de la Video.

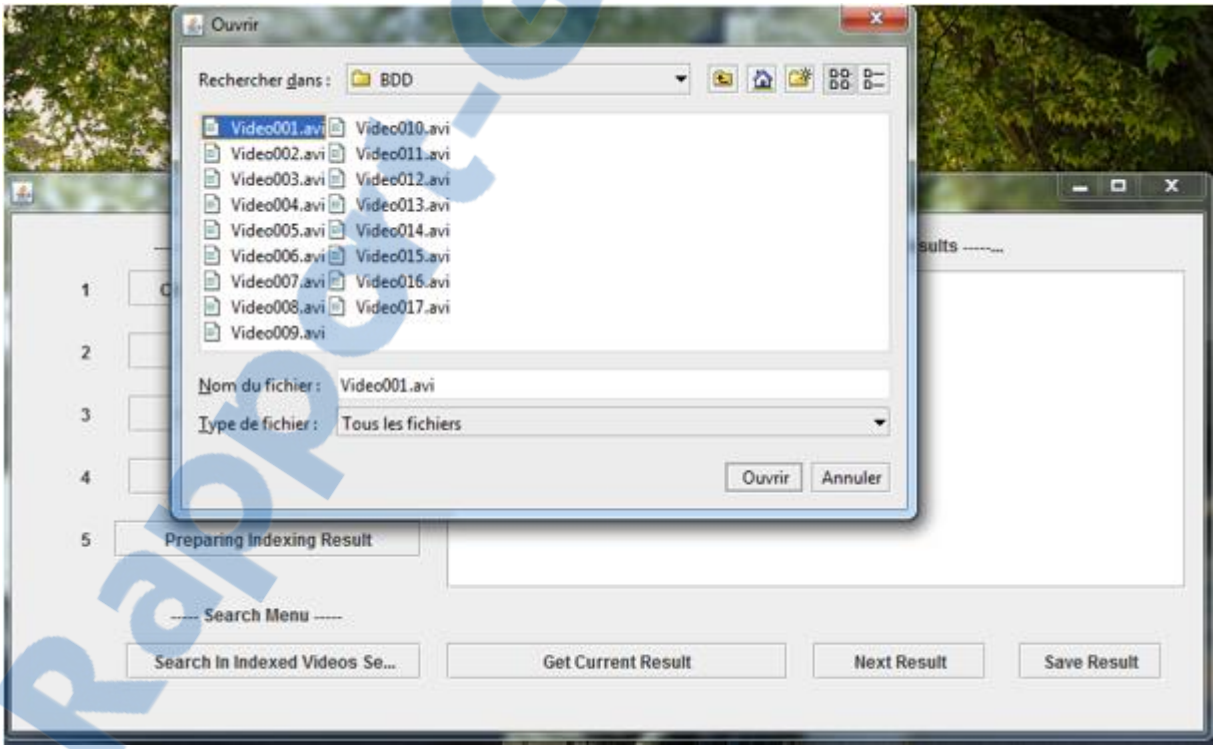


Figure 32. Choix de la vidéo à indexer.

L'étape suivante implique le module de bas niveau qui est exécuté à partir de l'interface d'indexation de la figure 33. Par contre, la figure 34 est une capture d'écran lors de l'exécution de ce module de bas niveau, où on peut voir les Blobs qui sont délimités en vert, englobés dans leurs Bounding Box et représentés en bleu. Une fois cette session est terminée, le résultat sera ajouté au fichier OWL de notre ontologie.

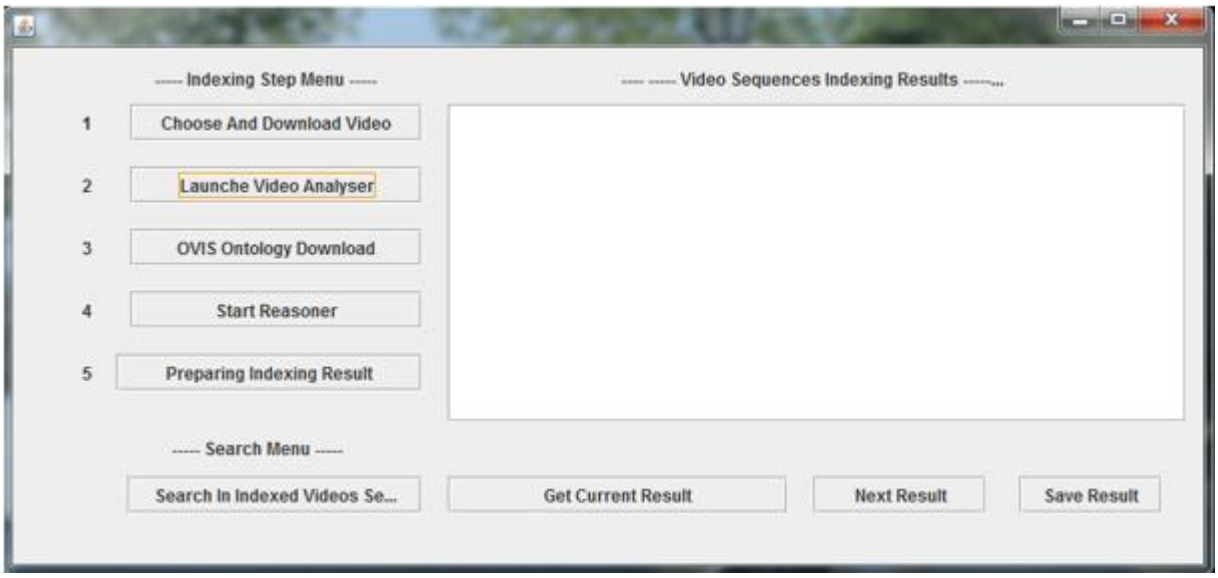


Figure 33. Sélection du bouton Launch Video Analyser.

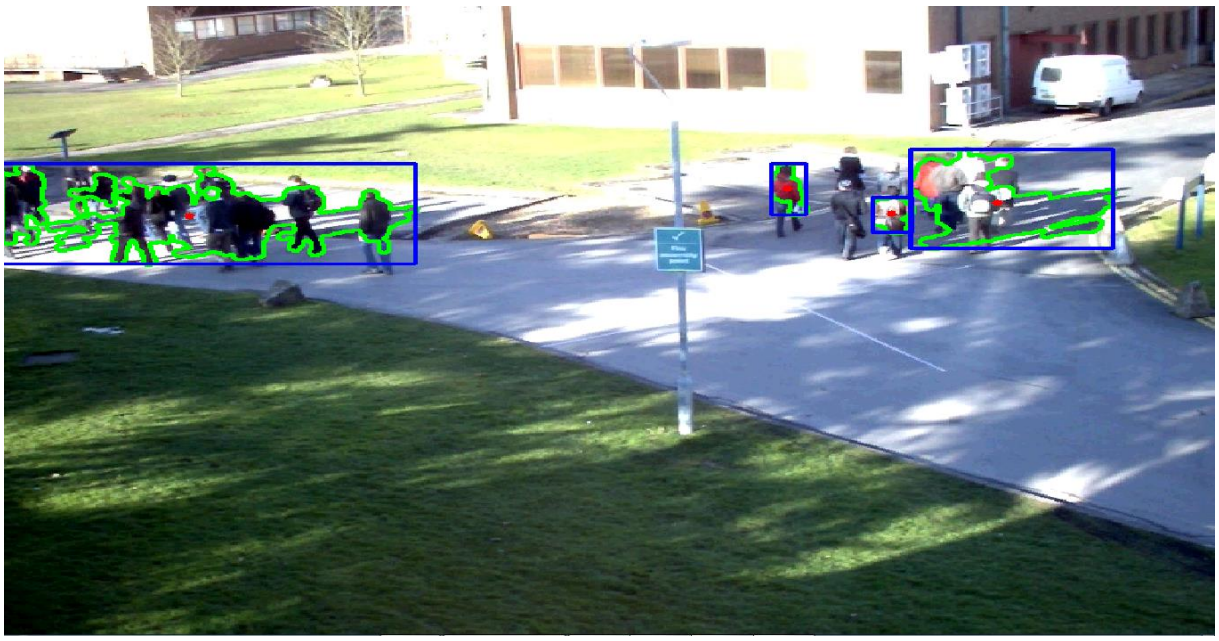


Figure 34. Capture d'écran lors de l'exécution du module de bas niveau.

L'étape suivante consiste à lancer l'ontologie renseignée avec les résultats du module de bas niveau (Individuals, Data-Properties, Object-Properties) comme c'est illustré dans les figures 35 et 36.

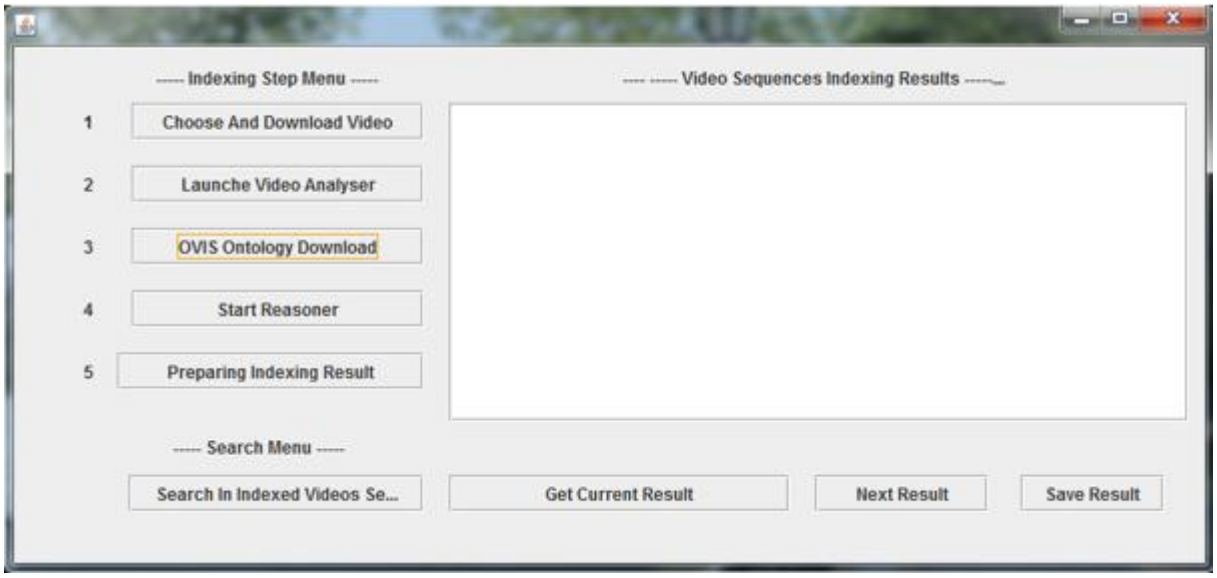


Figure 35. Sélection du bouton OVIS Ontology Download.

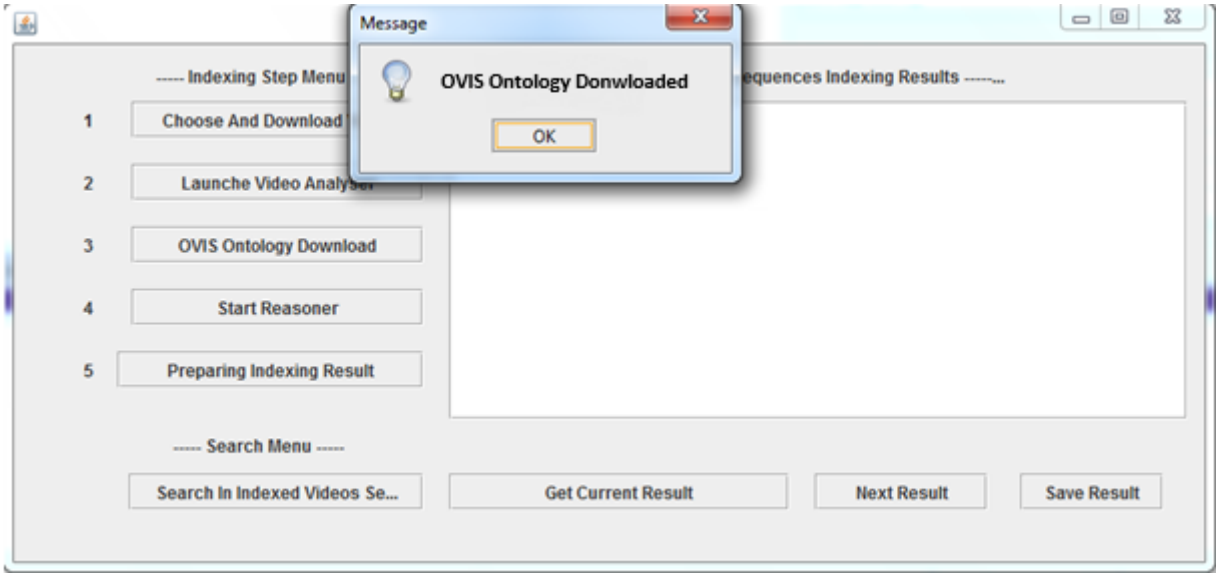


Figure 36. Message indiquant le chargement de notre.

Notre ontologie est exécutée avec l'éditeur Protégé 2000. La figure 37 illustre les différentes classes alors que la figure 38 montre les différents Individus qui peuplent notre ontologie.

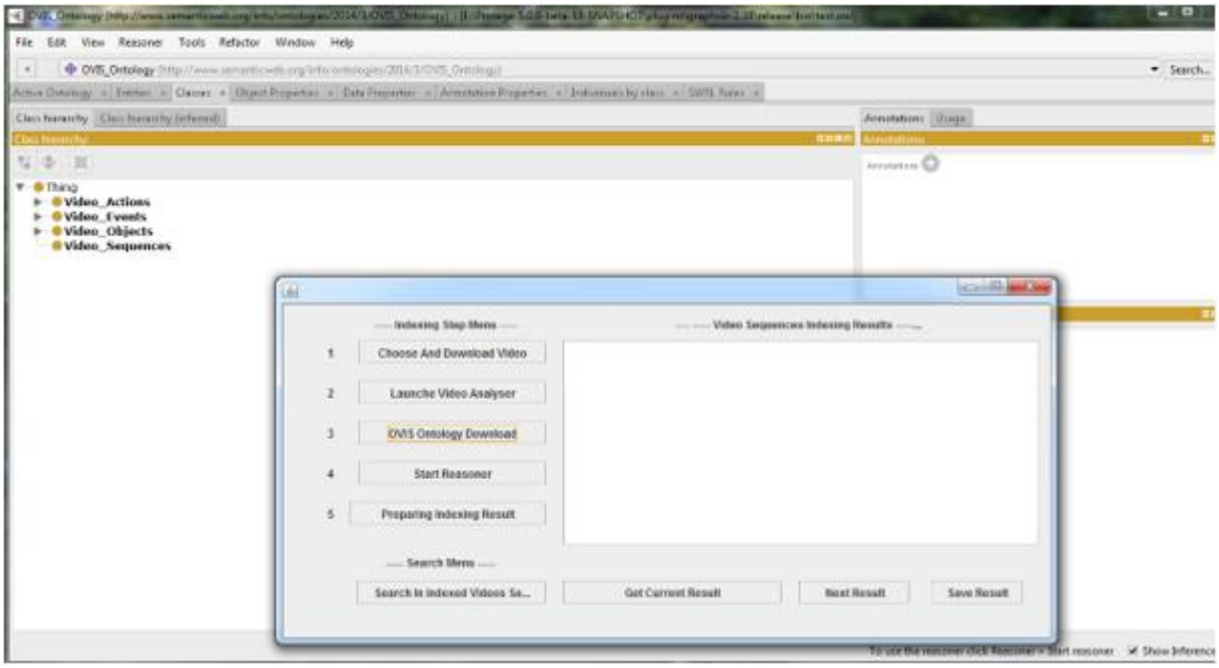


Figure 37. L'onglet Class hierarchy de notre ontologie.

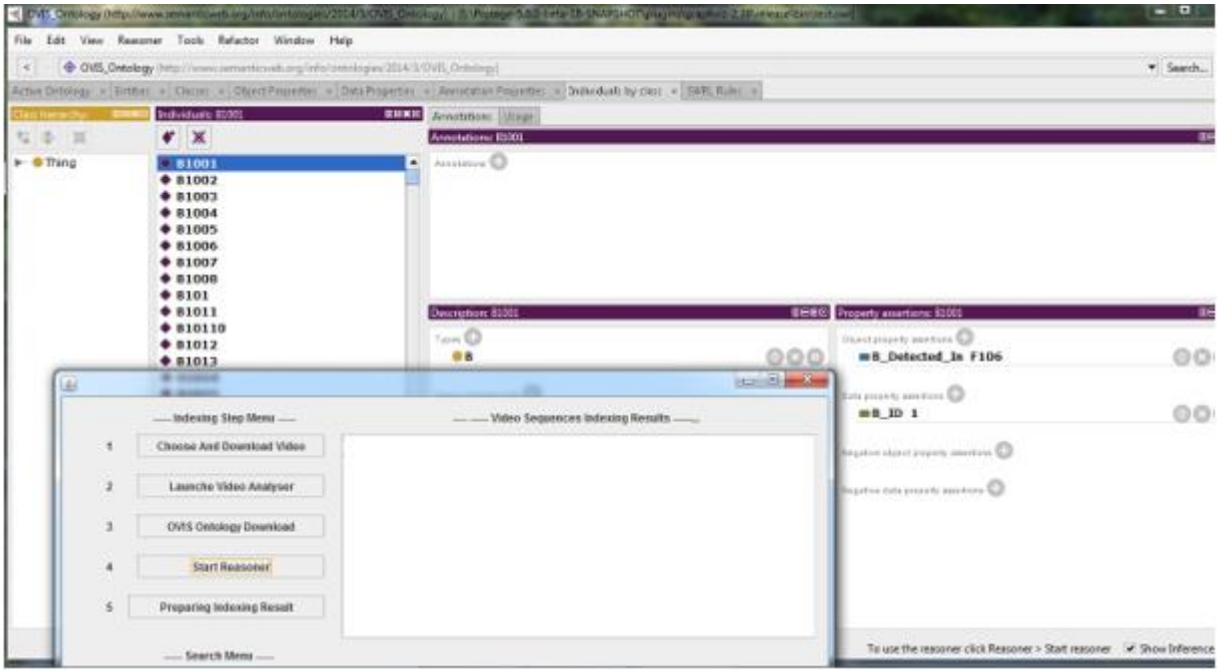


Figure 38. L'onglet Individual by class de notre ontologie.

Par ailleurs, la figure 38 illustre aussi la sélection du bouton Start Reasoner qui va exécuter les différentes règles SWRL. Une fois que cette exécution soit terminée, la figure 39 illustre l'affichage d'un message indiquant que le raisonneur Pellet termine l'exécution des règles SWRL.

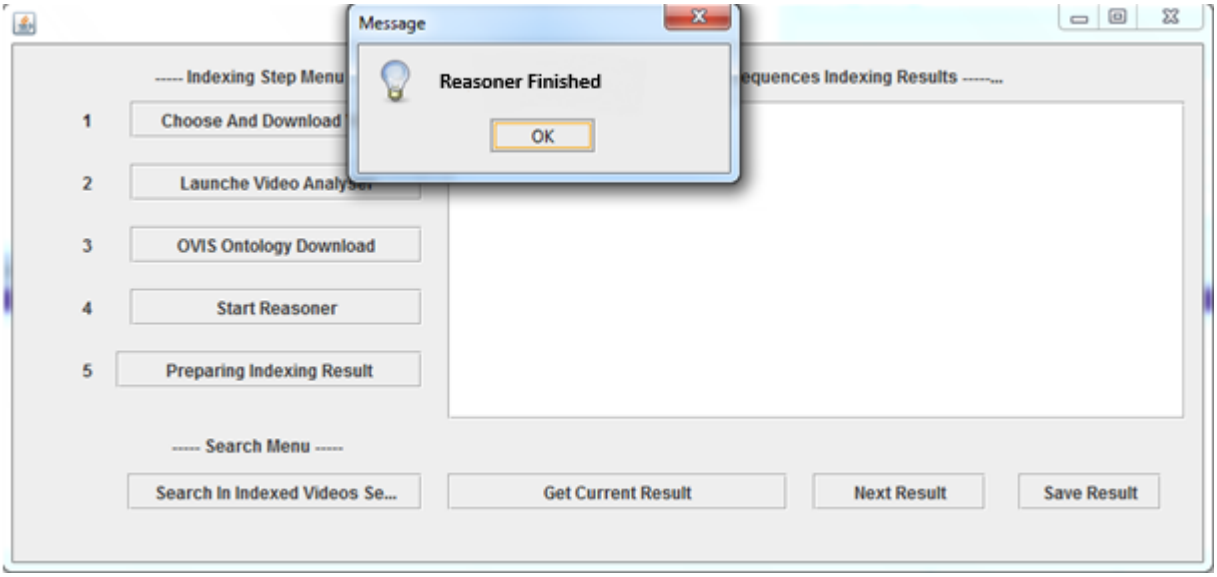


Figure 39. Message indiquant que le raisonneur Pellet termine l'exécution des règles SWRL.

Une fois le raisonneur termine l'exécution des différentes règles **SWRL**, les résultats d'inférence seront nombreux. On peut citer par exemple l'apparition de nouvelles informations comme les **Data-Properties et Object-Properties** concernant les nouveaux **Major Bounding Box**, les nouveaux **Groupe de personne** etc…. Cependant, ce qui nous intéresse le plus, ce sont les nouvelles connaissances en liaison avec la séquence vidéo qu'on a sélectionné pour indexer. Ces dernières concernent les évènements de cette vidéo ainsi que leur début et fin. Pour cela, une fois on sélectionne les bouton Get Current Result et Next Result, le système OVIS affichera les résultats d'inférence concernant ce type de donnée.

La figure 40 illustre la sélection du bouton Preparing Indexing Results, qui a pour but d'afficher les métas-données concernant la séquence vidéo, à savoir son nom et son emplacement.

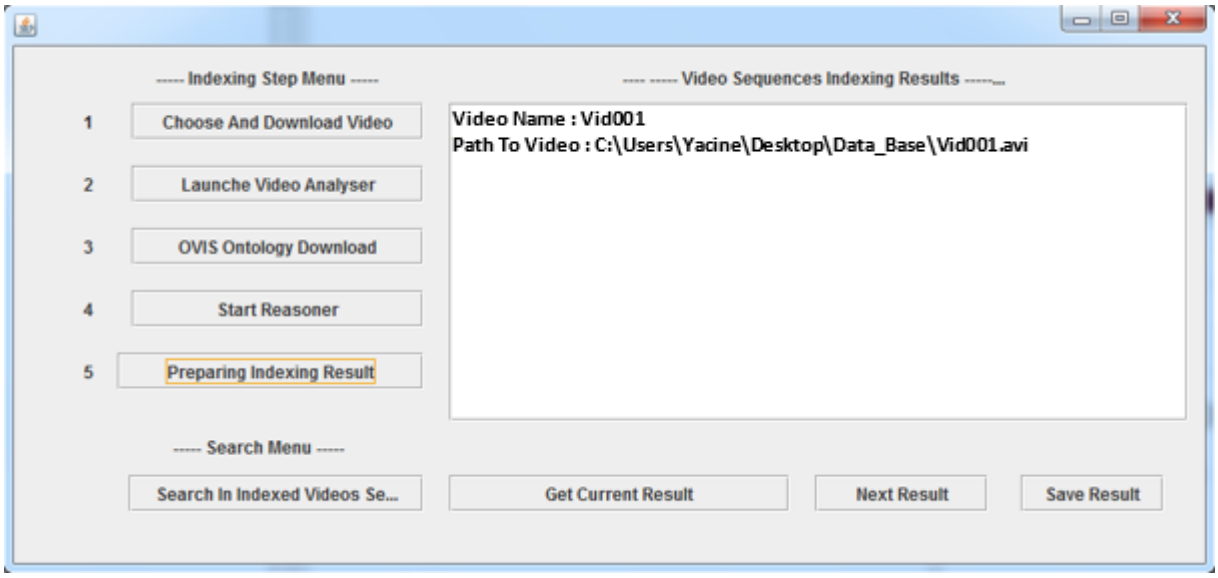


Figure 40. Sélection du bouton Preparing Indexing Results.

Les boutons des figures 41 et 42 affichent le résultat d'indexation des vidéos. Le bouton Get Current Result (Figure 41) affiche le premier type d'évènement alors que le bouton Next Result (Figure 42), affiche les autres types d'évènements.

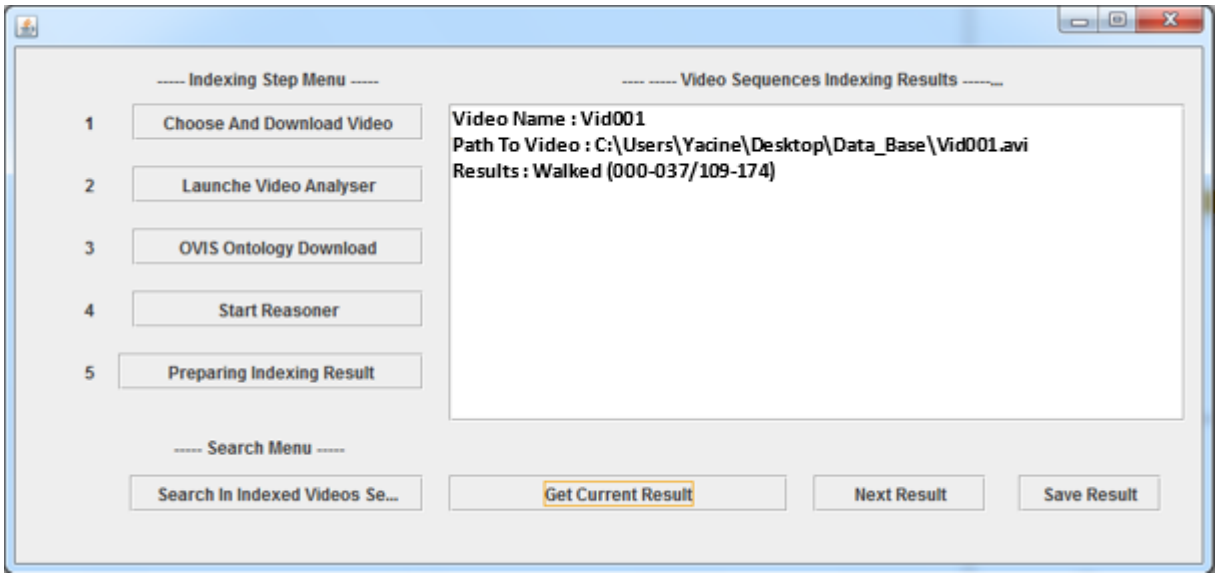


Figure 41. Sélection du bouton Get Current Result.

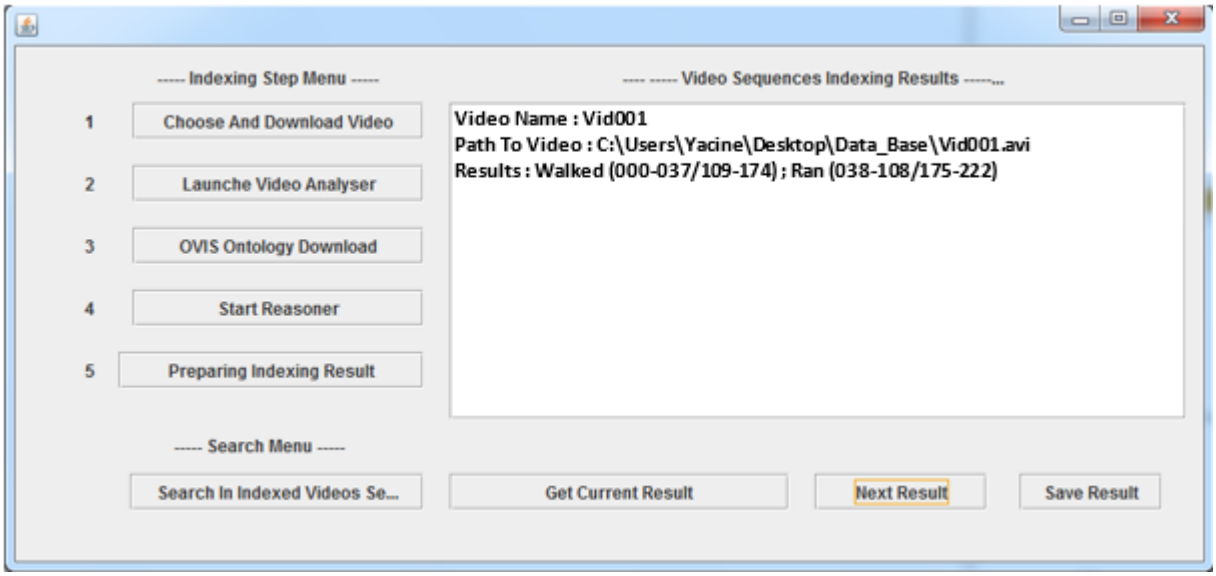


Figure 42. Sélection du bouton Next Result.

La sélection du bouton Save Result (figure 43) va créer un document texte (figure 44) dans le même emplacement disque que la vidéo à indexer et qui porte le même nom que cette vidéo. Ce document texte permet de sauvegarder les résultats d'indexation de cette vidéo comme l'illustre la figure 45.

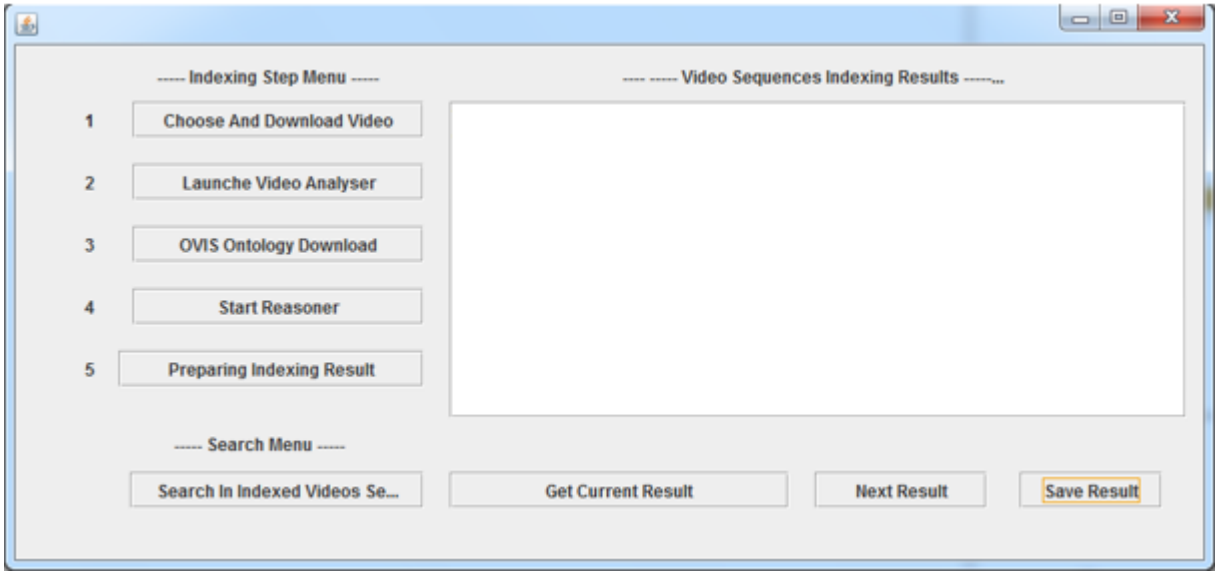


Figure 43. Sélection du bouton Save Result.

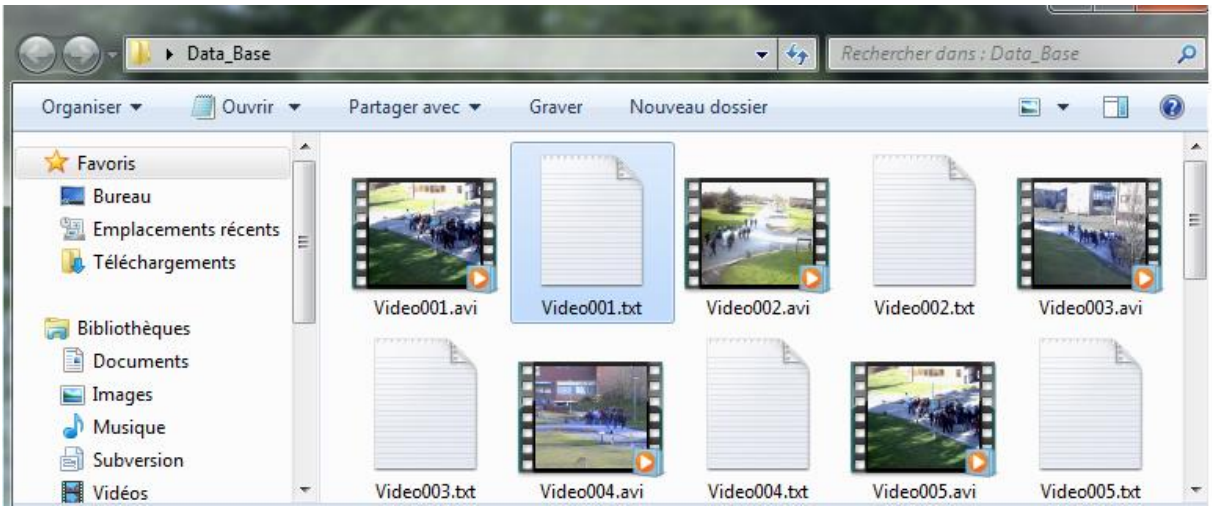


Figure 44. Création du document texte.

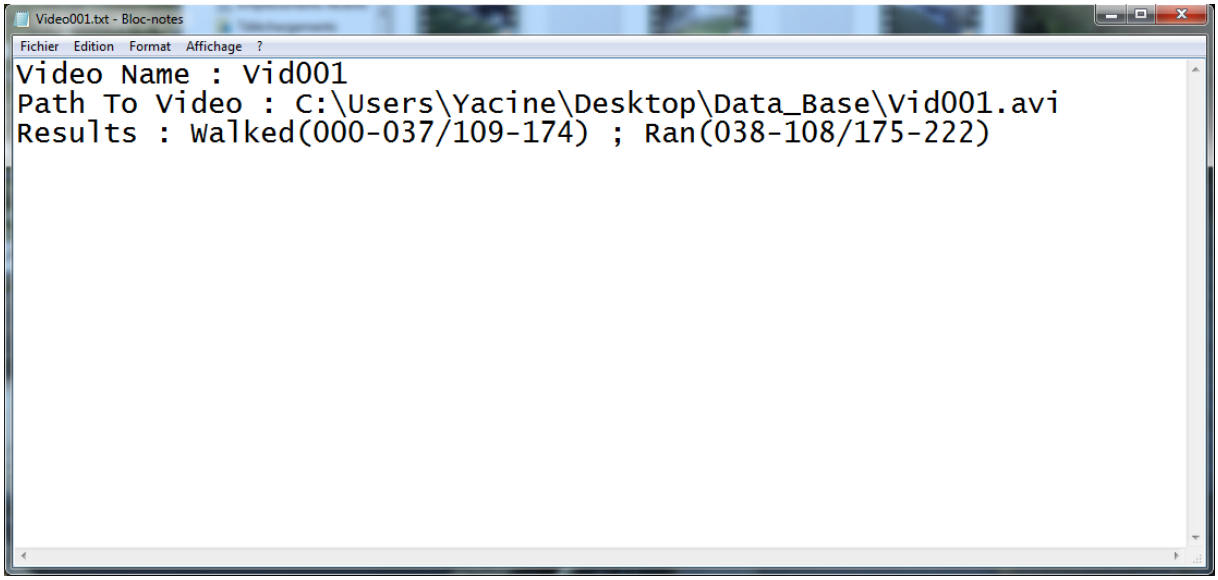


Figure 45. Exemple du contenu du document texte concernant la vidéo 001.

Lors de la phase de recherche, si on saisit par exemple le mot clé « Walked », le compilateur (parseur) de ces fichiers texte va les parcourir et rechercher ceux qui contiennent ce mot clé. Une fois qu'il le trouve dans le document texte, il affichera les informations écrites dans ce même document, sinon il passe au suivant, jusqu'à parcourir tous les documents texte disponibles.

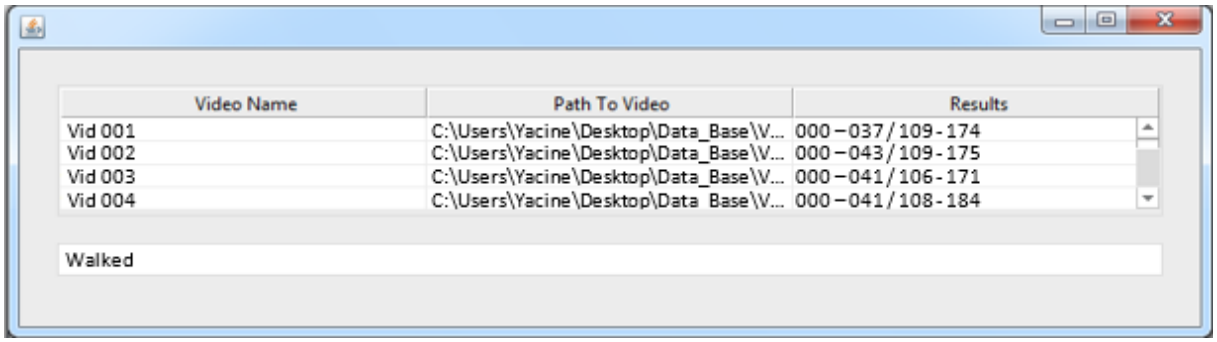


Figure 46. Recherche des évènements par notre système OVIS.

La figure 46 montre le résultat de recherche pour l'évènement "Walking" en montrant les différentes séquences vidéo qui contiennent cet évènement. Pour chaque séquence vidéo, le module de recherche affiche son nom, son acheminement ainsi que l'image du début et l'image de la fin de l'évènement.

Nos Publications

Communication dans une conférence Internationale avec comité de lecture

[1] M.Y. Kazi Tani, A. Ghomari and L. Kazi Tani, "*Toward a Hierarchical Based Video Indexing and Retrieval Approach*". In Proceedings of the International Conference on Information Communication Systems (ICICS'13), Irbid, Jordan, April 2013.

Communications dans des conférences Internationales avec comité de lecture (ISBN)

[2] M.Y. Kazi Tani, A. Ghomari and L. Kazi Tani, "*Hierarchical Video Indexing and Retrieval System*". Proceedings of the International Conference on Data Mining (DMIN'13) Robert Stahlbock Gary M. Weiss, ISBN: 1-60132- 239-9, Las Vegas, Nevada, USA, July 2013.

[3] M.Y. Kazi Tani, A. Ghomari, H. Belhadef, A. Lablack, I.M Bilasco, "*An ontology based approach for inferring multiple object events in surveillance domain*". Proceedings of the 2 nd IEEE Science and Information Conference (SAI'14) ISBN: ISBN: 978-0- 9893193-1- 7, August 17-29, London, UK 2014.

[4] Kazi Tani M.Y., Lablack A., Ghomari A. and Bilasco I. M., "*Events Detection Using a Video-Surveillance Ontology and a Rule-Based Approach*". 1st International Workshop on Computer vision + ONTology Applied Cross-disciplinary Technologies in Conjunction with ECCV 2014. Lecture Notes in Computer Science LNCS 8926, 2015, pp. 299-308.

[5] Kazi Tani M.Y., Ghomari A, Dali Youcef Lamia., Lablack A. and Bilasco I. M., "*An Audio Indexing and Retrieval Approach using a Video Surveillance Ontology*". Proceedings of the 5nd IEEE Science and Information Conference (SAI'17), 18-20 July 2017, London, UK 2017.

Article de revue internationale avec comité de lecture

[1] Kazi Tani M.Y., Ghomari A, Lablack A. and Bilasco I. M., OVIS: ontology video surveillance indexing and retrieval system". International Journal of Multimedia Information Retrieval (Springer), Vol 6, Issue 4, 2017.

Références

[1] Alberto Del Bimbo, Pietro Pala, and Enrico Vicario. Spatial ar- rangement of color flows for video retrieval. In IEEE International Conference on Multimedia and Expo (ICME), pages 413–416, 2001.

[2] Bertrand Chupeau and Ronan Forest. An evaluation of the effectiveness of color attributes for video indexing. In SPIE Storage and Retrieval for Media Databases, pages 470–481, 2001.

[3] Juan M. Sanchez, Xavier Binefa, Jordi Vitria, and Petia Radeva. Linking visual cues and semantic terms under specific digital video domains. Journal of Visual Languages and Computing, 11(3):253–271, 2000.

[4] Yi Wu, Yueting Zhuang, and Yunhe Pan. Content-based video retrieval integrating human perception. In SPIE Storage and Retrieval for Media Databases, pages 562–570, 2001.

[5] Stamatia Dasiopoulou, Vasileios Mezaris, Ioannis Kompatsiaris, V.-K. Papastathis, and Michael G. Strintzis. Knowledge assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology, (10):1210–1224, Oct 2005.

[6] Kazi Tani MY, Ghomari A, Lablack A, Bilasco IM: OVIS: Ontology Video Surveillance Indexing and Retrieval System. International Journal of Multimedia Information Retrieval, Volume 6, Issue 4, PP 295-316, December 2017.

[7] M. OConnor, H. Knublauch, S. Tu, B. Grosof, M. Dean, W. Grosso and M. Musen, (2005) "Supporting rule system interoperability on the semantic web with swrl," in 4th International Semantic Web Conference (ISWC), pp. 974-986.

[8] Patrice Maubourguet, editor. *Grand Larousse Universel*. Larousse, 1991.

[9] Daniel Lecomte, Daniel Cohen, Philippe De Bellefonds, and Jean Barda. *Les normes et les standards du multimédia, 2eme édition.* Dunod, 2000.

[10] Sharon Flanck. Multimedia Technology in Context. *IEEE Multimedia*, Juillet - Septembre 2002.

[11] John Edwards and Linda Dailey Paulson. Smart Graphics: A New Approach to Meeting User Needs. *Computer*, 35(5):18–21, Mai 2002

[12] Peter Hanappe. Design and implementation of an integrated environment for music composition and synthesis. PhD thesis, Universit´e Paris 6, Avril 1999.

[13] Mbarek CHARHAD. Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique. Thèse de doctorat, Université Joseph Fourier, Novembre 2005.

[14] Sturm, P. (1997). Vision 3D non calibrée : contributions à la reconstruction projective et étude des mouvements critiques pour l'auto-calibrage (Doctoral dissertation, Institut National Polytechnique de Grenoble-INPG).

[15] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M.,Hafner, J., Lee, D., Petrovic, D., Steele, D., Yanker, P.: Query by Image and Video Content: the QBIC System. *IEEE Computer* 28(9), pp. 23-30, 1995.

[16] Babaguchi N., Kawai Y., Kitahashi T., Event Based Video Indexing by Intermodal Collaboration, Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99), Orlando, FL, USA, pp. 1-9, 1999

[17] Zhang T., Tretter D., An Overview of Video Abstraction Techniques, HP Laboratory Technical Report HPL-2001-191, 2001

[18] Kankahalli M. S, Chua T.-S., Video Modeling Using Strata-Based Annotation, IEEEMultimedia, 7(1), pp. 68-74, Mar 2000.

[19] SALWAY, A. *Video Annotation: the role of specialist text*. PhD Dissertation, Dept. of Computing, University of Surrey, 1999.
URL: http://www.mcs.surrey.ac.uk/Personal/A.Salway/publications.html.

[20] Corridoni J. M., A. Del Bimbo, D. Lucarella, and H. Wenxue: "Multiperspective navigation of movies", Journal of Visual Languages and Computing, 7:445-466, 1996.

[21] Hampapur, A., Jain, R., and Weymouth, T., "Digital Video Segmentation", Proc. ACM Multimedia 94, San Francisco, CA, pp. 357-364, 1994.

[22] D. Kless, L. Jansen, J. Lindenthal and J Wiebensohn, (2012) "A method for reengineering a thesaurus into an ontology," Frontiers in Artificial Intelligence and Applications, pp. 133-146.

[23] A. Badii, C. Lallah, M. Zhu and M. Crouch, (2009) "The dream framework: Using a network of scalable ontologies for intelligent indexing and retrieval of visual content," in Proceeding of the International Conference on Web Intelligence and Intelligent Agent Technology, pp. 551-554.

[24] M. Rodrguez-Muro and D. Calvanese, (2012) "High performance query answering over dl-lite ontologies," in Proceeding of the International Conference on Principles of Knowledge Representation and Reasoning (KR), pp. 308-318.

[25] A. Scherp, C. Saathoff, T. Franz and S. Staab, (2011) "Designing core ontologies," Journal Applied of ontology, vol 03, pp. 177-221.

[26] R. Benmokhtar and B. Huet, (2011) "An ontology-based evidential framework for video indexing using high-level multimodal fusion," Multimedia Tools and Applications, vol. 55, no. 3, pp. 1–27.

[27] A. Rector, S. Brandt, N. Drummond, M. Horridge, C. Pulestin and R. Stevens, (2012) "Engineering use cases for modular development of ontologies in owl," Journal Applied of ontology, vol. 02, pp. 113-132.

[28] B. Smith and Werner Ceusters, (2010) "Ontological realism as a methodology for coordinated evolution of scientific ontologies," Journal Applied of ontology, vol.03, no. 4, pp. 139-188.

[29] S. Kara, Z. Alan, O. Sabuncu, S. Akpnar, N. K. Cicekli and F. N. Alpaslan, (2012) "An ontology-based retrieval system using semantic indexing," Information Systems Journal, vol. 04, pp. 294-305.

[30] T. Mossakowski, C. Lange and O. Kutz, (2013) "Three semantics for the core of the distributed ontology language," in Proceeding of the International Joint Conferences on Artificial Intelligence (IJCAI), pp. 3027-3031.

[31] L. Ballan, M. Bertini, A. D. Bimbo and G. Serra, (2010) "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies," Multimedia Tools and Applications, vol. 02, pp. 313-337.

[32] A. D. Bagdanov, M. Bertini, A. D. Bimbo, G. Serra and C. Torniai, (2007) "Semantic annotation and retrieval of video events using multimedia ontologies," in Proceeding of the International Conference on Semantic Computing (ICSC), pp. 713-720.

[33] M. Bertini, A. D. Bimbo, C. Torniai, C. Grana and R. Cucchiara, (2007) "Dynamic pictorial ontologies for video digital libraries annotation," in Proceeding of the 1st ACM Workshop on The Many Faces of Multimedia Semantics, pp. 47-56.

[34] M. Xue, S. Zheng and C. Zhang, (2012) "Ontology-based surveillance video archive and retrieval system," in Proceeding of the 5th International Conference on Advanced Computational Intelligence (ICACI), pp. 84-89.

[35] J. Lee, M. H. Abualkibash and P. K. Ramalingam, (2008) "Ontology based shot indexing for video surveillance system," in Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, pp. 237-242.

[36] L. Snidaro, M. Belluz and G. L. Foresti, (2007) "Representing and recognizing complex events in surveillance applications," in Proceeding of the 4th IEEE International Conference Advanced Video and Signal based Surveillance (AVSS), pp. 493-498.

[37] L. Calavia, C. Baladrn, J. M. Aguiar, B. Carro and A. Sanchez-Esguevillas, (2012) "A semantic autonomous video surveillance system for dense camera networks in smart cities," sensors, pp. 10407-10429.

[38] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris and M. G. Strintzis, (2007) "Ontology-driven semantic video analysis using visual information objects," in Proceedings of the semantic and digital media technologies second international conference on Semantic Multimedia, pp. 56-69.

[39] S. Saad, D. D. Beul, M. Said and M. Pierre, (2012) "An ontology for video human movement representation based on benesh notation," in Proceeding of the IEEE International Conference on Multimedia Computing and Systems (ICMCS), pp. 77-82.

[40] I. Trochidis, E. Tambouris and K. Tarabanis, (2007) "An ontology for modeling life-events" in Proceeding of the IEEE International Conference on Services Computing (SCC), pp. 19-20.

[41] W. Bohlken and B. Neumann, (2009) "Generation of rules from ontologies for high-level scene interpretation," Lecture Notes in Computer Science, pp. 93-107.

[42] R. Nevatia, J. Hobbs and B. Bolles, (2004) "An ontology for video event representation," in Proceeding of Computer Vision and Pattern Recognition, pp. 119-128.

[43] L. Bai, S. Lao, W. Zhang, G. J. F. Jones and A. F. Smeaton, (2008) "Video semantic content analysis framework based on ontology combined mpeg-7," Lecture Notes in Computer Science, pp. 237-250.

[44] J. C. SanMiguel, J. M. Martinez and A. Garcia, (2009) "An ontology for event detection and its application in surveillance video," in IEEE Proceeding of AVSS, pp. 220-225.

[45] A. Utasi, A. Kissand T. Sziranyi: (2009) "Statistical filters for crowd image analysis". In Performance Evaluation of Tracking and Surveillance workshop,at CVPR, Miami, Florida ,pp. 95–100.

[46] Antoni B. Chan, Mulloy Morrowand Nuno Vasconcelos: (2009) "Analysis of crowded scenes using holistic properties". In 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS).

[47] TRECVID. TRECVID 2016 challenge [online]: http://www nlpir.nist.gov/projects/tv2016/tv2016.html

[48] F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, T. Mironidis, V. Kaltsa, A. Ioannidou, S. Symeonidis, K. Avgerinakis, S. Andreadis, I. Gialampoukidis, S. Vrochidis, A. Briassouli, V. Mezaris, I. Kompatsiaris, I. Patras. (2016) "ITI-CERTH" at TRECVID.

[49] Z. Zhao, M. Wang, R. Xiang, S. Zhao, K. Zhou, M. liu, S. He, Y. Zhu, Y. Zhao, F. Su. (2016) "BUPT-MCPRL" at TRECVID.

[50] Kazi Tani MY, Ghomari A, Lablack A, Bilasco IM (2015) Events detection using a video-surveillance ontology and a rule-based approach. In Computer vision + ONTology applied cross-disciplinary technologies workshop (CONTACT) in conjunction with European conference in computer vision (ECCV), pp 299–308.

[51] Kazi Tani MY, Ghomari A, Belhadef H, Lablack A, Bilasco IM (2014) An ontology based approach for inferring multiple object events in surveillance domain. In: IEEE science and information conference (SAI), pp 404–409

[52] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, (5-6):907–928, Nov-Dec 1995.

[53] Thomas R. Gruber. *Ontolingua: A mechanism to support portable ontologies*. KnowledgeSystems Laboratory Technical Report KSL-91-66, Stanford University, version 3.0, CA, 1992.
URL: http://www-ksl.stanford.edu/knowledgesharing/papers/index.html#ontolingua-long

[54] NicolaGuarino. *Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration.*In Information Extraction: A Multidisciplinary Approach to anEmerging Information Technology, SCIE 1997, M. T. Pazienza (Eds.), Springer Verlag, pp. 139-170,1997.
URL: http://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html

[55] Natalya Fridman Noy and Carole D. Hafner. The State of the Art in ontology Design: A survey and Comparative Review. AI Magazine, 18(3): 53-74 (Fall 1997).

[56] Mike Uschold and Robert Jasper. A Framework for Understanding and Classifying Ontology Applications. KRR5-99, Stockholm, Sweden, 1999.

[57] Kifer M., Lausen G. et Wu J. Logical Foundations of Objetc-Oriented and Frame- Based Languages. Journal of the ACM, 1995.

[58] M. Chein, M. L. Mugnier. Conceptual Graphs: Fundamental Notions. Revue d'Intelligence Artificielle, vol. 6, n. 4, p. 365-406, 1992.

[59] Gomez Pérez A., Benjamins V.R. (1999) "Overview of Knowledge Sharing and Reuse Components: Ontologies and problem-Solving Methods". Proceeding of the IJCAI-99 workshop on Ontologies and problem-SolvingMethods (KRR5), Stockholm (Suède).

[60] Nathalie Hernandez. Ontologies de domaine pour la modélisation du contexte en recherche d'information. Thèse de doctorat, décembre 2005.

[61] Roussey Cathrine. Une méthode d'indexation sémantique adaptée aux corpus multilingues, Thèse de doctorat, décembre 2001.

[62] Sowa J., Conceptual Structures: Information Processing in Mind and Machine, AddisonWesley, 1984.

[63] Roussey Cathrine. Une méthode d'indexation sémantique adaptée aux corpus multilingues, Thèse de doctorat, décembre 2001.

[64] Michel Klein. Change Management for Distributed Ontologies. PhdThesis. Vrije Universiteit Amsterdam, URL : http://www.cs.vu.nl/~mcaklein/thesis/

[65] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau et John Cowan. *Extensible Markup Language (XML) 1.1*. http://www.w3.org/TR/2004/RECxml11-20040204/ (en ligne au 16 juin 2005).

[66] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau et John Cowan. *Extensible Markup Language (XML) 1.1*. http://www.w3.org/TR/2004/RECxml11-20040204/ (en ligne au 16 juin 2005).

[67] Thi-Lan Le.,Indexation et Recherche de vidéo pour la vidéosurveillance. Thèse de doctorat, Université de Nice-Sophia Antipolis 2009

[68] P. Kuznetsova, V. Ordonez, T. Berg and Y. Choi, (2014) "Treetalk: Composition and compression of trees for image descriptions," TACL: Transactions of the Association for Computational Linguistics, pp. 351-362.

[69] R. Socher, A. Karpathy, V. Q. Le, C. D. Manning and A. Y. Ng, (2014) "Grounded compositional semantics for finding and describing images with sentences," TACL: Transactions of the Association for Computational Linguistics, pp. 207-218.

[70] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, (2014) "Show and tell: A neural image caption generator," CoRR: Computing Research Repository, abs/1411.4555.

[71] R. Kiros, R. Salakhutdinov and R. S. Zemel, (2014) "Unifying visual-semantic embeddings with multimodal neural language models," CoRR: Computing Research Repository, abs/1411.2539.

[72] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, (2014) "Explain images with multimodal recurrent neural networks," CoRR : Computing Research Repository, abs/1410.1090.

[73] PETS. Pets 2012 challenge [online]: http://www.cvg.reading.ac.uk/PETS2012/a.html.

[74] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle and A. Courville, (2015) "Describing videos by exploiting temporal structure," CoRR : Computing Research Repository, abs/1502.08029.

[75] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal and B. Schiele, (2014) "Coherent multi-sentence video description with variable level of detail," in German Conference on Pattern Recognition (GCPR).

[76] M. Rohrbach, W. Qiu, I. Titov, T. Stefan, M. Pinkal and B. Schiele, (2013) "Translating video content to natural language descriptions," in IEEE International Conference on Computer Vision (ICCV).

[77] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney and K. Saenko, (2014) "Translating videos to natural language using deep recurrent neural networks". CoRR: Computing Research Repository, abs/1412.4729.

[78] MULLER, Pierre-Alain et GAERTNER, Nathalie. Modélisation objet avec UML. Paris : Eyrolles, 2000.

[79] OpenCV. The OpenCV API [online]: http://docs.opencv.org/3.1.0/index.html

[80] Protege. The protege projectt [online]: http://protege.stanford.edu.

[81] Sirin E. B. Parsia B. Cuenca Grau B. Kalyanpur A. Katz Y. (2003) "Pellet: A Practical OWL-DL Reasoner". Journal of Web Semantics.

[82] Allen, J. F. (1983) "Maintaining knowledge about temporal intervals". Communications of the ACM, 26, pp. 832–843.

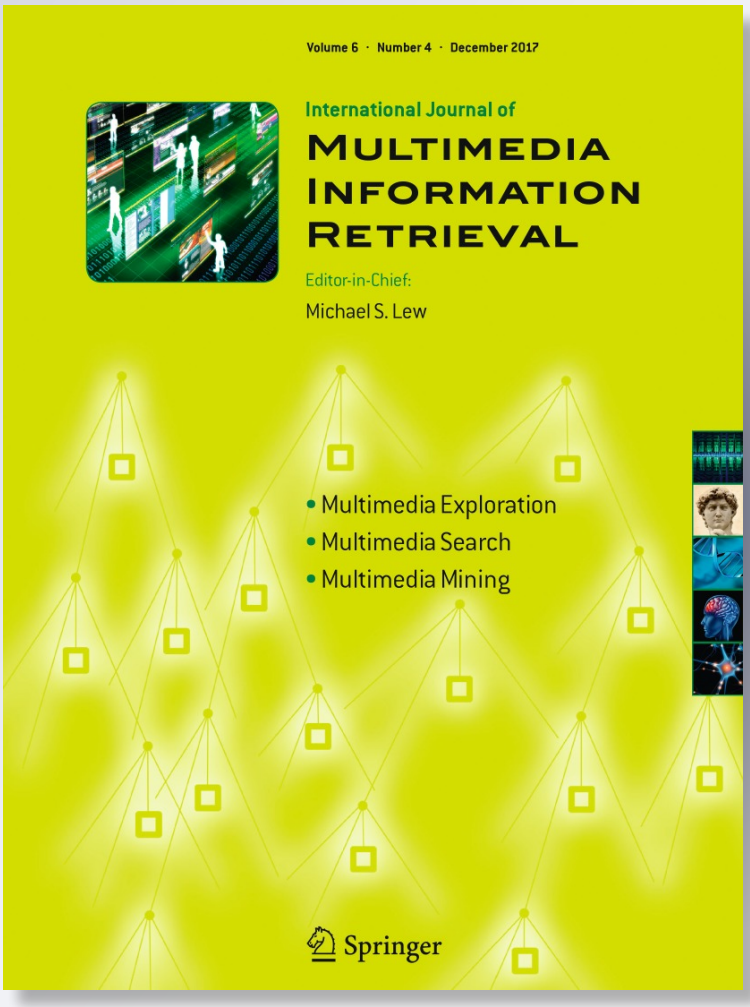
OVIS: ontology video surveillance indexing and retrieval system

Mohammed Yassine Kazi Tani, Abdelghani Ghomari, Adel Lablack & Ioan Marius Bilasco

International Journal of Multimedia Information Retrieval

ISSN 2192-6611
Volume 6
Number 4

Int J Multimed Info Retr (2017)
6:295-316
DOI 10.1007/s13735-017-0133-z



Springer

Int J Multimed Info Retr (2017) 6:295–316
DOI 10.1007/s13735-017-0133-z

REGULAR PAPER

OVIS: ontology video surveillance indexing and retrieval system

Mohammed Yassine Kazi Tani[1] · **Abdelghani Ghomari**[1] · **Adel Lablack**[2] · **Ioan Marius Bilasco**[2]

Received: 27 April 2017 / Revised: 10 August 2017 / Accepted: 5 September 2017 / Published online: 18 September 2017

Abstract Nowadays, the diversity and large deployment of video recorders result in a large volume of video data, whose effective use requires a video indexing process. However, this process generates a major problem consisting in the semantic gap between the extracted low-level features and the ground truth. The ontology paradigm provides a promising solution to overcome this problem. However, no naming syntax convention has been followed in the concept creation step, which constitutes another problem. In this paper, we have considered these two issues and have developed a full video surveillance ontology following a formal naming syntax convention and semantics that addresses queries of both academic research and industrial applications. In addition, we propose an ontology video surveillance indexing and retrieval system (OVIS) using a set of semantic web rule language (SWRL) rules that bridges the semantic gap problem. Currently, the existing indexing systems are essentially based on low-level features and the ontology paradigm is used only to support this process with representing surveillance domain. In this paper, we developed the OVIS system based on the SWRL rules and the experiments prove that our approach leads to promising results on the top video evaluation benchmarks and also shows new directions for future developments.

Keywords Video surveillance ontology · Video indexing · Crowdsourced events · Semantic gap · Naming syntax convention · OVIS system · SWRL rules

1 Introduction

During the last few years, the semantic multimedia indexing process becomes a major research topic in computer vision and machine learning, due to the huge increase in the size of recorded data and the diversity of application domains like sport, broadcasting, news, cooking and surveillance. The need to find an effective tool to index and store this large volume of data for future uses must be satisfied. Therefore, many scientists explore new ways to improve the existing approaches or to develop new ideas in the video indexing domain. Currently, there are two dominant categories of indexing approaches. The first one consists of using low-level features in an automatic system where the second one uses metadata or keywords in a manual high-level system. Even if the combination of those two approaches could offer an efficient indexing system, it does not overcome completely the semantic gap problem.

Generally, the semantic gap outlines the differences between the video sequence information perceived by human experts and the interpretation of the results obtained from low-level analyzers. Several works used an ontology-based approach to handle this problem. For instance, Kless et al.

✉ Mohammed Yassine Kazi Tani
yassine.kazi@gmail.com

Abdelghani Ghomari
ghomari65@yahoo.fr

Adel Lablack
adel.lablack@univ-lille1.fr

Ioan Marius Bilasco
marius.bilasco@univ-lille1.fr

[1] RIIR Laboratory, Computer Science Department, Exact Sciences and Applied Faculty, University of Oran 1 Ahmed Ben Bella, Oran, Algeria

[2] Research Center in Signal Informatics and Automatic of Lille (CRIStAL), University of Lille 1, Lille, France

[1] present a thesaurus or taxonomies as the best method for the creation of ontology. Atta et al. [2] develop a framework based on a network of scalable ontologies that index a large repository of special effects video clips. This proposed framework enables intelligent retrieval for the film post-production domain. Mariano et al. [3] realize a system to answer ontological queries that include many specific optimizations. On one hand, they exploit the ABox (assertion component) skills that generally represent the assertion component or instances of the class. On the other hand, they respect the TBox (terminological component) properties that generally describe a system in terms of controlled vocabulary. Scherp et al. [4] propose the notion of a core ontology, like a system based on the logical notion of reducibility, rather than on the distinction between generic and domain ontologies. Benmokhtar et al. [5] use an ontology paradigm integrated with a neural network approach for detecting a concept purpose. Rector et al. [6] consider annotation using existing ontologies as a good practice. Smith et al. [7] present a specific theory of ontology called Ontological Realism to build high-quality ontologies, using both philosophical views (i.e., the study of the existing entities and the way that they are related to each other) and computer science ones (i.e., the conceptualization of a domain). Hernandez-Leal et al. [8] propose the use of ontologies as a tool to reduce the semantic gap between low and high-level information, and present the foundations of ontology to be used in an intelligent video surveillance system.

Our main contributions in this paper are as follows:

- To our knowledge, this is the first work that introduce the idea of following a unified ontology structure to formalize both the objects and the actions as well as the events in surveillance domain. We have also modeled the non-sequential relationship between events using Allen's interval algebra.
- Secondly, we extend the video surveillance domain representation by considering new concepts characterizing events in the industrial domain, whereas the approaches proposed in the literature focused only on academic aspects.
- Our ontology is more complete compared with those proposed in the literature and gives a large coverage of important objects and events in the surveillance domain.
- Fourth, this is the first work that proposes a diversity of applications, which are not limited to indexing purposes, but also applies to scene description, benchmark creation, etc.
- Finally, we contribute the use of the proposed ontology in a rule-based indexing and retrieval approach, by generating SWRL rules. These rules are deployed in the middle and high-level step of an event detection process, that are supplied with low-level descriptors, rather than using the classical descriptors and classifiers in all-event detection steps.

The remainder of this paper is organized as follows: Sect. 2 focuses on related works. In Sect. 3, we present our ontology with its naming syntax convention and semantics. Section 4 describes a comparative study with other ontologies. Section 5 presents the various domains of application of our ontology. The OVIS system architecture is illustrated in Sect. 6. Preliminary of the video surveillance ontology indexing and retrieval system developed here are given in Sect. 7. The final section provides some concluding remarks.

2 Related works and background

In video surveillance applications, an ontology could be used in the indexing process to support the detection of an event such as an abnormal behavior, crowd situations of people or traffic monitoring. Indeed, the description of a domain covered by the ontologies and the reasoning results that are generated increase the accuracy of the indexing process. Several approaches have been proposed using an ontology.

Soner et al. [9] use an ontology to extract instances from a document corpus, and add them to the knowledge basis. Till et al. [10] handle the problem of using a variety of languages and propose a distributed ontology language DOL that allows to use its own preferred ontology formalism considering the interoperability with the others. Ballan et al. [11] recognize events in broadcast news and video surveillance domains by embedding knowledge into the ontology. Bagdanov et al. [12] use a multimedia ontology that contains visual prototypes representing each cluster that acts as a bridge between the domain and the video structure ontology. They present a system that gives a solution to the semantic gap between the high-level concepts and low-level descriptors. Bertini et al. [13] classify the events and the objects that are observed in video sequences by adding new instances of visual concepts to their ontology through updating mechanisms of the existing concepts. This approach used in both generic and specific domain descriptors attempts to identify visual prototypes that represent elements of visual concepts. To overcome the problem of manual rules creation by a human expert, Bertini et al. [14] proposed an adaptation of the First Order Inductive Learner (FOIL) technique for Semantic Web Rule Language (SWRL) [15], called FOILS. Xue et al. [16] propose an ontology-based surveillance video archive and retrieval system. Lee et al. [17] classify and index video surveillance streams through the creation of the framework called Video Ontology System (VOS). Snidaro et al. [18] use a set of rules in SWRL language for event detection purposes in the video surveillance domain.

The problem that arises in the use of an ontology paradigm to support the indexing process is to find the best way for the creation of this ontology. Moreover, some previous works have used the ontology tool and demonstrated their efficiency in helping and managing the indexing and retrieval process. However, they have based their experimental studies on events that consider only one or two relevant objects in a video clip. In the contextual cases, they consider events such as an abandoned or stolen object; whereas in the moving cases, they consider events like a person who walks from right to left, an airplane flying, and so on. The problem that arises is how to ensure the efficiency of the ontology in the indexing and retrieval process when the user requires multiple object events or crowd sourced events considering a set of relevant objects (e.g., queries about a regular group of people walking, a group of people running and a group of people splitting).

Calavia et al. [19] developed an intelligent video surveillance ontology system that analyzes objects movements and identify abnormal and alarming situations. However, the domain application covered by the documentation is not consistent with the ontology representation. Papadopoulos et al. [20] proposed a genetic algorithm for optimizing the size of each ontology element (e.g., concepts). In this way, they consider the variable relevant importance of global and local information to detect the different ontology elements. Nevertheless, the relationship named "some/some" is used instead of "all/some". Sawsan et al. [21] constructed a video movement ontology for automatic annotation of human movement's purposes in the classic Benesh notation. However, it is not clear whether their ontology is formal or not, and there is something wrong in the use of the "Is-A relation" in a non-transitive way as the relation between the two concepts (i.e., media and video) in multimedia representation part. Trochidis et al. [22] proposed a well-structured ontology approach to model life events described as a graph of connections between concepts with representing a particular domain. Nevertheless, this approach has some limitations about mainstream ontology and its application in video analysis. Bohlken et al. [23] considered the problem of high-level scene interpretation suggesting a novel architecture based on the generation of rules from an OWL-DL ontology. However, the concept of vehicle entering a zone is not conceptual, because it represents an action between vehicles and zone concepts.

Nevatia et al. [24,25] developed two languages called VERL (Video Event Representation Language) and VEML (Video Event Markup Language), for describing an ontology of events, and annotating instances of the events, respectively. However, a confusion occurs between the object language that describes the referent in the subject domain and the meta-language that defines this object language. Bai et al. [26] presented a video semantic content analysis framework based on ontology. A high-level concept is described referring to this domain of application and combined with the MPEG-7 standards for expressing low-level content analysis algorithms. Nevertheless, this ontology confuses between the relation "Is-A", and "Instance-Of". For example, the combination of many algorithm instances with the "Is-A" relation in an ontology replaces the "Instance-Of" relation. SanMiguel et al. [27] proposed an ontology-based approach to represent the prior knowledge of a video event analysis consisting of two types of knowledge: the application domain and the analysis system. The domains knowledge involves all the high-level semantic concepts (objects, events, context, etc.), while the system knowledge includes the abilities of the analysis system (algorithms, reactions to events, etc.). However, this ontology determines only the best visual analysis framework (or processing scheme), and does not handle the inference for object tracking and event detection.

The detection process is often represented using three levels: low-level, middle-level and high or semantic one. Different approaches could be found in the literature using descriptors in the low and middle level and classifiers in the semantic one. Utasi et al. [28] proposed a statistical descriptor approach detecting three kinds of events: regular activity, running and splitting. This approach consists in using a background extraction technique followed by calculating the optical flow of foreground pixels. However, it does not detect many events like walking and formation. Chan et al. [29] used a model based on global properties to detect events such as walking, running, splitting, formation and local dispersion. Their approach characterizes the crowd flow using a dynamic texture. However, this model does not process overlapping that could occur between events.

Other semantic based approaches participate on TRECVID Surveillance Event Detection (SED) task 2016. Markatopoulou et al. [30] proposed a system for surveillance event detection based on fisher vector encoding method and SVM models to learn how to separate each activity from the others. However, this approach detects many false alarms. Zhao et al. [31] used different approaches to detect surveillance events, and their overall system consists of two parts: the retrospective part and the interactive part. The retrospective part implements pedestrian detection, pedestrian tracking and event detection. The interactive part determines the events after fixing the false and missing rate. However, this method considered a limited number of events.

After a deep analysis of all the problems noted above, we introduce an innovative approach in this work, by creating an ontology and implementing the OVIS indexing and retrieval system that considers all the above observations. In this paper, we have reconsidered our previous works

[32,33] where we presented only our SWRL rules based approach allowing to handle a video surveillance ontology to detect a single or multiple objects events. In the present work, we have improved and extended our previous approach by implementing a complete video surveillance ontology with a very precise step creation syntax. This extension describes numerous objects and events that can appear in a video surveillance domain. Furthermore, the creation of our ontology is more complete considering new concepts that characterize events in the industrial domain. Moreover, we have not based our approach only on indexing purposes, but also in scene/video description and benchmark creation domains. Finally, we used our ontology to prove the efficiency of our approach. We have generated SWRL rules for event detection, and used them in both middle level and high or semantic level, rather than using the classical descriptors and classifiers in all-event detection steps. These SWRL rules use results that are supplied by low-level descriptors for event detection purposes. We have also tested the performance of these rules by experimenting videos from the PETS 2012 challenge [34] and SED task from TRECVID challenge [35]. The PETS challenge represents multiple view sequences that handle different crowd activities and contain multiple objects events (e.g., group walking and group splitting). It allows providing the existence of each event, in these sequences, with start/end as well as transitions between these different events. The main goal of TRECVID (The TREC Video Retrieval Evaluation) is to promote progress in content-based analysis and retrieval from digital video via open metrics-based evaluation. The Surveillance Event Detection (SED) task focuses on developing new approaches able to detect observations of different events. It consists in a subset of 10 h of videos recorded using multi-camera derived from Gatwick airport. Seven events are identified: PersonRuns, CellToEar, ObjectPut, PeopleMeet, Embrace, PeopleSplitUp, and Pointing. In our work, we focus on three events: PersonRuns (Running), PeopleMeet (Formation) and PeopleSplitUp (Splitting).

3 Ontology hierarchy description

Our ontology creation based on a naming syntax convention includes most of the existing concepts appearing in a video surveillance domain. Indeed, our approach is an improvement of SanMiguel et al. [27] work, where we used the same modeling that defines the high-level relationships between objects and events to compose single and multiple object activities. We believe that this modeling is the most suitable to represent the video surveillance domain. However, regarding the different levels that compose ontologies, we use this modeling only in the high level expressed as level 2 "L2" in Tables 1 and 3 below. We also introduced new concepts representing events used in the industry like intrusion.

3.1 Semantics of our ontology

Generally, the semantic interpretation of video sequences is the critical step in an indexing process. It corresponds to the translation of the low-level features extracted from the visual analysis module into the video sequence meaning. Here, we used an ontology paradigm as a tool characterizing a video surveillance system. Our semantic ontology is represented by different interacting concepts where each concept carries one or more properties as a "Data_Property" described in detail below.

3.1.1 Ontology concepts

The concepts of the ontology proposed here correspond to the categorization of the video surveillance domain, regarded as generalization/specialization relationships. In order to have a complete representation of all the objects or events that can happen in a video surveillance domain, we formally divide the ontology into four main categories concepts representing Video_Actions, Video_Events, Video_Objects and Video_Sequences.

Figures 1 and 2 below show how the four main categories of concepts are linked together, where each one forms an interconnection with the other. First, all the available video sequences in the video database must be indexed with one or more concepts that appear in the Video Events category. A relationship exists between Video Events and Video Actions since an event is a composition of one or more actions. Furthermore, another connection exists between Video Actions and the video objects category. Actually, scene description is formed by the video object components category where the actions occur. Finally, video sequences category encloses objects that belong to the video objects category. In the following, we describe each category.

Video objects The different kinds of objects represent the principal entities that interact in the video sequences. The video_objects category involves all the objects that can appear in a video surveillance scene. A large variety of objects interacts with each other to create a video surveillance action. According to their mobility skills, the objects can be assembled into two main categories. The Contextual Objects having no mobility skills and the Mobile Objects with mobility skills. Eventually, we can add a third category representing an image region of interest (ROI), characterized by Low-Level Features. It represents all the data extracted from the visual analysis module. Table 1 below illustrates these three categories dividing the video objects into five levels (from L1 to L5).

Table 1 Video_Objects hierarchy

L 1	L 2	L 3	L 4	L 5
Video_ Objects	Contextual_ Objects	Fixed_Objects	Human_Creation _Objects	Air-Conditioner/Building/Electrical-Pole/Equipment/Floor/Panel/Parking-Lot/Road/Stairs/Stairs-Barrier/Wall/Zone/Glass Barrier
			Natural_ Objects	
		Portable_Objects	General_Using	Box/Chair/Door/Plant /Reception-Desk/Surveillance-Camera/Table/Window/Curtain/Sofa
			Self_Using	Cellphone/Document/Luggage (Bag, Suitcase)
	Mobile_Objects	Airplane		
		Boat		
		Train	Long-Distance-Train/City-Tramway/Underground	
		Animal		
		Human	Person/Group-Of-Person	
		Ground_Traffic	Bicycle	
			Ground-Vehicle	Bus/Car/Truck
			Motorcycle	
	Low_Level_ features	Bounding-Box (BB)		
		Frame		
		Major_Bounding-Box (MBB)		
		Temporary_Bounding-Box (TBB)		
		Temporary_Group-Of-Person (TGP)		
		Blocks(B)		

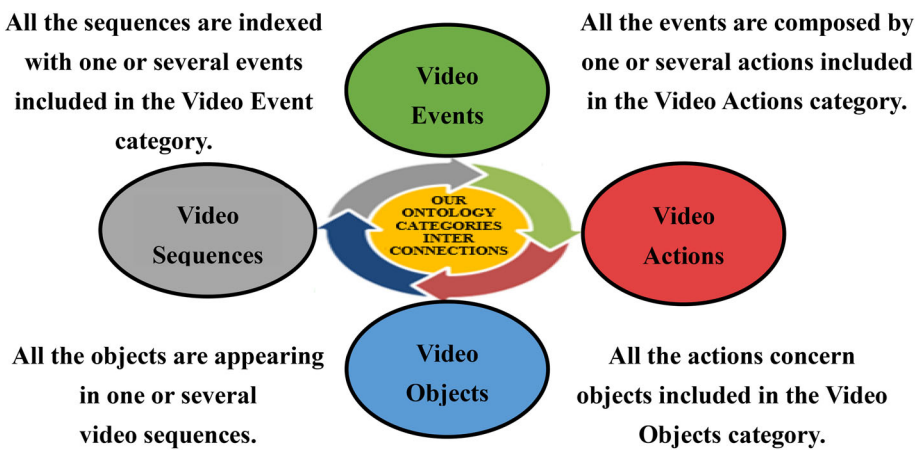


Fig. 1 Interconnection between the four main categories of concepts in the proposed ontology

- The category of Contextual Objects is divided into two subclasses representing Fixed Objects (that can never be moved by other objects) and Portable Objects (that have the possibility of moving with other objects). Fixed Objects correspond to those of human creation (e.g., air conditioners and panels) and natural objects (e.g., grass and land). Portable Objects represent all general and self-using objects (e.g., a box, a chair and a cellphone).
- The category of Mobile Objects encloses four subclasses that have the ability of self-moving, such as animals, airplanes, trains, boats, humans and traffic. Traffic corresponds to bicycles and motorcycles. The Human subclass

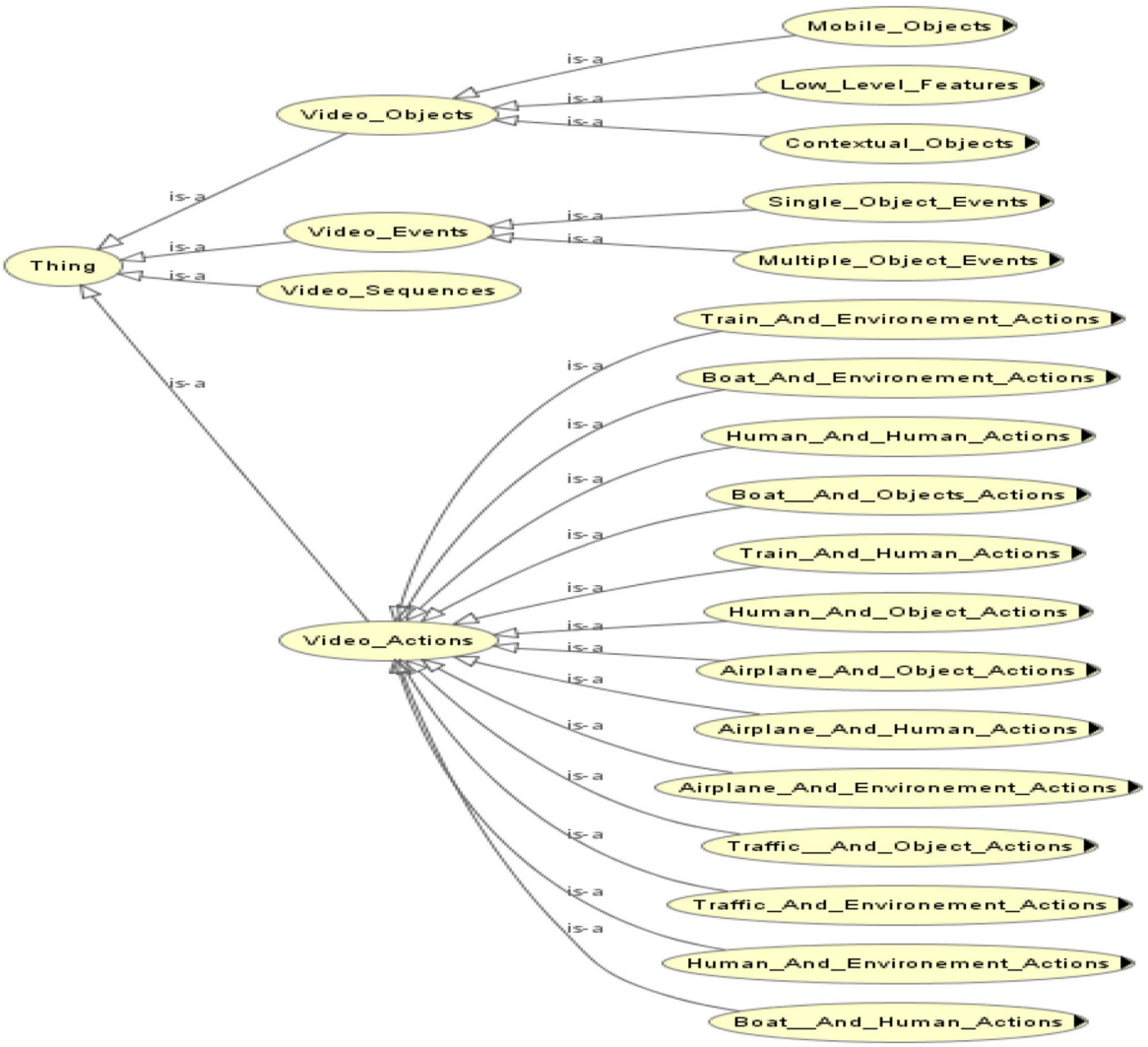


Fig. 2 Illustration of the "Is-a" relation representing the four main categories of our ontology

corresponds to a human entity like a person or Group-Of-Person.

- The category of Low-Level Features includes all the results of the low-level analysis module that could be used to help the indexing process for event detection purposes, such as blocks, bounding box and frame.

Video actions In video surveillance domain, the action concept represents the behavior of different objects detected in the video sequence in a time frame window. This category includes the actions that can be expected in video surveillance events. Therefore, several varieties of occurring objects can produce multiple kinds of actions. Generally, we can expect five categories of objects that belong to Video Objects category: airplane, boat, train, traffic (road traffic) and humans. In our ontology, we divide these actions into different subclasses according to the nature of the different objects interactions:

- Interactions with environment: such as human walking, human stopping and airplane landing.
- Interactions with humans: human attacking, human meeting, etc.
- Interactions with objects: such as a human breaking an object.

As described in Sect. 3.2, Table 2 below presents a summary of the different subclasses of the Video Actions parts of our ontology, illustrated at different levels (L1 to L3). We separate all the actions according to their degree of priority. The first subclass degree of the Video Action category expressed in level 2 represents the action actors and properties with which they interact (e.g., Human_And_Objects_Actions). The second subclass degree expressed in level three corresponds to the actions themselves. Among them, we note those that are widely reported in the literature such as (Split,

Table 2 Video_Actions hierarchy

L 1	L 2	L 3
Video_Actions	Train_And_Environment_Actions	Fire_detected/Arrival_detected/Stationed
	Train_And_Human_Actions	Upped/ Downed/Crossed_Forbidden_Zone
	Boat_And_Environment_Actions	Navigated
	Boat _And_Human_Actions	Threw_bag
	Boat _And_Objects_Actions	Approached_bank/Drived_away_the_bank
	Airplane_And_Environment_Actions	Crashed_Off/Flew/Landed/Took_Off
	Airplane_And_Human_Actions	Disembarked/Embarked
	Airplane_And_Objects_Actions	Registered_Luggage/Took_Luggage
	Human_And_Objects_Actions	Downed_Stairs/Read _Document/Broke_Object/Browsed_Object/ Counted_Ground-Vehicule/ Counted_Human/ Crossed_Virtual-Line/Left_Luggage/Put_Object/ Removed_Luggage/Rested_On_Chair/ Smoked_Cigarette/Upped_Stairs
	Human_And_Human_Actions	Attacked/Chased/Evacuated/ Fell_Down/Flow_Opposed/Formed/Fought/Helped/ Hit/Local_Dispersed/Met/Split/Stole/Talked/Waited
	Human_And_Environment_Actions	Appeared/Counted_Speed/Disappeared/ Entered_Area/Face_Recognized/Fell/Intruded/ Left_Area/Loitered/Overcrowded/Ran/ Skateboarded/Slipped/Stopped/Trespassed/Walked
	Traffic_And_Object_Actions	Crashed_Object
	Traffic_And_Environment_Actions	Parked

Met and Ran) and those related to the industrial domain. Thus, for example walking and running events represent generally a group-of-person that interacts with the environment by multiple walked and ran actions. Intrusion event related to industrial domain represents the attempt of a person to enter a restricted zone. Considering new concepts that characterize events in the industrial domain, is one of the objectives of our proposal work contribution.

Video events The video event concept represents a composition and succession of one or several actions appearing in a video sequence. In the present ontology, the Video_Events category encloses all the different events that could happen in a video stream. Each event representing the formation of actions encloses one or several relevant objects that interact with each other.

Table 3 describes the four levels (L1–L4) representing the different subclasses of the Video Events parts in our ontology. Each level corresponds to a degree of priority, as described in Sect. 3.2. The first degree is related to the number of relevant objects and divides our Video Events category into two main subclasses representing, respectively, single and multiple object events. The second degree is related to the nature of objects represented by seven types: Group-Of-Person, Multiple_Ground-Vehicle, Airplane, Train, Boat, Person and Single_Ground-Vehicle. The final degree corresponds to the interaction between these objects and the other concepts such as human, environment or objects.

Video sequences The video sequences category represents the class of all the videos indexed by the OVIS system and the instances represent the Video Database.

After presenting the Ontology concepts part, in the following we describe DataProperty and ObjectProperty parts.

3.1.2 Ontology DataProperty

The Data_Property represents the real information related to individual's concepts. In our ontology, DataProperty includes all the properties related to one or more concepts. Table 4 displays the DataProperty hierarchy divided into three levels (L1–L3). The top level is split into seven subclasses related to the types of DataProperty, such as Event properties and Frame Properties. Each of them enclosing one or more data properties like Event_Place (if the event represents indoor or outdoor events), Number_Frame, etc.

3.1.3 Ontology ObjectProperty

The Object_Property concerns the concepts of the ontology interactions and is divided into three levels as shown in Table 5. A complete representation of all interactions

Table 3 Video_Events Hierarchy

L 1	L 2	L 3	L 4
Video_Events	Multiple_Objects_Events	Group-Of-Person_Events	Interaction_Group-Of-Person_And_Environment Interaction_Group-Of-Person_And_Human
		Multiple_Ground-Vehicle_Events	Interaction_Multiple_Ground-vehicle_And_Objects
	Single_Objects_Events	Air-Plane_Events	Interaction_Airplane_And_Environment Interaction_Airplane_And_Human Interaction_Airplane_And_Objects
		Train_Events	Interaction_Train_And_Environment Interaction_Train_And_Human
		Boat_Events	Interaction_Boat_And_Environment Interaction_Boat_And_Human Interaction_Boat_And_Objects
		Person_Events	Interaction_Person_And_Environment Interaction_Person_And_Human Interaction_Person_And_Objects
		Single_Ground-Vehicle_Events	Interaction_Single_Ground-Vehicle_And_Environment

Table 4 Top_Data_Property Hierarchy

L 1	L 2	L 3
Top_Data_Property	Event_Properties	Event_Place/Nature_Event
	Detected_Objects	Bottom_Left_Point_X/Bottom_Right_Point_Y/ Detected_In_Frame/ Direction/Ended_F/Height/ID/Leaving_Object_Way/Major_BB/ MBB_True/Number/Number_Of_Person/Posture/Speed/Started_F/ Top_Left_Point_X/Top_Right_Point_Y/Weight/etc.
	Entering_Exit	
	Frame_Properties	Number_Frame/Number_BB_In_Frame/Number_MBB_In_Frame/ Started_MBB/etc.
	Type	
	Time	
	Video_Sequence_Properties	Video_URI/Number_Of_Frame/Started_F_Formed_Event/ Ended_F_Formed_Event/Took_Place_Before_R/Started_R/etc.

Table 5 Top_Object_Property Hierarchy

L 1	L 2	L 3
Top_Object_Property	Human_Against_Human	Asked_Direction/Chased/Formed_Final_Meta_Group/Attacked/ Formed_With/Had_Diff_End_Position/Had_Different_Direction/ Had_Same_Start_Position/Had_Started_Meta_Group/Helped/Hit/ Met_With/Pushed/Split_With/Spoke_With/Stole/Walked_With/etc.
	Human_Against_Objects	Walked_Around/Attempted_To_Open/Browsed_On/Downed/Left/ Loitered_In/Occurred_In/Put/Rested_On/Stood_Near/Upped/etc.
	Objects_Against_Objects	Detected_In/Crashed_With/Flew_In/Landed_In/Parked_In/ Represented/Took_Off_From/etc.

between our ontology concepts is obtained by subdividing the top level into three subclasses reflecting the interaction between the objects. We can consider two categories of objects like Humans and Objects (referring to subcategories of Video Objects other than humans); while, the interactions represent Human_Against_Human, Human_Against_Objects, and Objects_Against_Objects. In the last level, each type of interaction encloses its Object-Property, such as Asked_Direction, Walked_Around, Detected_In. For describing some examples of DataProperty and ObjectProperty related to objects, Fig. 3 presents properties of Group-Of-Person and Bounding-Box Objects.

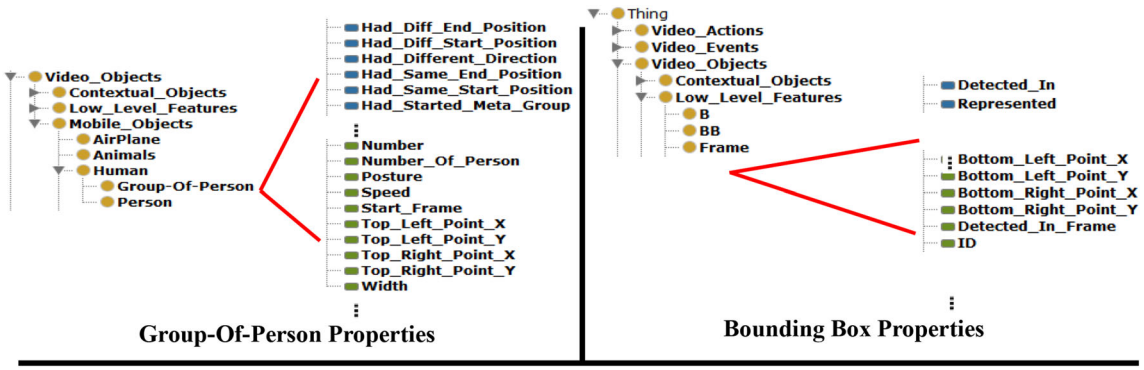


Fig. 3 Group-of-person and bounding box properties illustration

3.2 Syntax of our ontology

We have created a naming syntax convention composed of several rules to obtain a complete and consistent ontology. First, all the new concepts created in the Video_Event Category must be named in the same way as the previous ones in this category. Each concept is composed of three parts:

- The interaction (each concept starts with an interaction name).
- The name of objects existing in the Video Object category.
- The property with which this object interacts.

Furthermore, all the concepts of our ontology are generalized, and their details are classified in the form of Object_Properties and Data_Properties. As a typical example, the notion of time, posture, position and interaction are represented as Object_Properties or Data_Properties, and not under the subevent class. Moreover, the concept naming duplication must be avoided and Multiple_Event separated from Single_Event, Human_Actions from Objects_Actions, as well as Person Events from Objects Events. In addition, the Data_Properties and the Object_Properties must be generalized and the duplication avoided. For instance, instead of having a Data_Property called Name-Of-Animals for the concept animal and another one called Name-Of-Building for the concept building, we introduce a generalization and call the DataProperty "Name," using it for both concepts. The events and actions must be separated according to their degree of priority. Concerning the events, the first degree is attributed to multiple or single object, the second degree is linked to the nature of the object (Group-Of-person, Person and Ground-Vehicle.), and the third one represents the object interaction (with Human, with Environment or with Object). As far as the actions are concerned, the first degree is set according to the action actors and the concepts with which they interact, while the second degree is set with the action itself.

In the present work, the Group-Of-Person Object is regarded as the formation of two or more individuals. To be more concise, each word in the formation of a concept, an Object_Property or a Data_Property, must start with a capital letter. Relationships are determined between concepts (ObjectProperty) in the three categories: Object-Against-Object, Human-Against-Object, and Human-Against-Human. The new concepts added to the Video Actions Category are made of one word representing the action, except for the category object, where we need to specify the interaction of our concepts (Human, Ground, Vehicle, Airplane). The nature of interaction with an object in the category Video_Actions must be specified if the action does not consider all types of objects. Moreover, all the concepts of our ontology are created by unifying the words with an underscore (_), except for the composed objects, linked with a hyphen (-).

4 Comparison with other ontologies

Table 6 shows a comparison of the present ontology with the others in literature. Each version has advantages and weaknesses related with the domain representation covering, consistency, etc. As far as our ontology is concerned, it is essentially based on avoiding all these negative points, as explained in Sect. 3. The aim of this work is to develop a strong and efficient ontology for applications in various domains as discussed in Sect. 5.

Our ontology uses a naming syntax convention and semantics for consistency, formalism, conceptualization and sufficiently clear relationships between concepts. Moreover, we create a complete video surveillance ontology that includes most of the events arising from both the research and industrial domains. Furthermore, our ontology handles the small domain representation covering problem. Finally, our ontology has been developed for future usages like inference rules

Table 6 Comparative study between our ontology and other ones

Metrics points ontologies	Consistent	Formal	Large domain representation covering	Conceptual	Separate IS-A from INSTANCE-OF relations	Use inference rules (SWRL)
Our Ontology	OK	OK	OK	OK	OK	OK
Calavia et al. [19]		OK	OK		OK	OK
Sawsan et al. [21]	OK		OK	OK		OK
Trochidis et al. [22]	OK	OK		OK	OK	OK
Bohlken et al. [23]	OK		OK		OK	OK
Bai et al. [26]		OK	OK	OK		OK
SanMiguel et al. [27]	OK	OK		OK	OK	

(SWRL) which enhance their efficiency with new knowledge in event detection as described in Sects. 6 and 7.

5 Case studies in various domains

We have proposed a complete and consistent ontology that covers major video surveillance events. This complete knowledge representation could be useful in various domains. In the following subsections, we present some interesting application domains that could use the proposed ontology.

5.1 Benchmark creation

A benchmark represents challenges accepted and practiced by the scientific community to solve problems in various domains. In video surveillance, most of the benchmarks like PETS and TRECVID handle events detection. The formalism of our ontology could help these benchmarks in the selection process of the events appropriately.

5.2 Scene description

Scene description represents all the objects that act/appear in the scene. These objects form either background objects (objects acting for a long time in the scene) and/or foreground objects (new objects appearing in the scene).

Recently, a particular attention is given to the process of describing images automatically. Kuznetsova et al. [36] proposed to consider the task of image description as a retrieval problem, and create a hand-designed approach able to describe images in a wild field. It is based on retrieving similar captioned images from a large database, before generating new description by generalizing and recomposing the retrieved captions. This approach involves typically an intermediate generalization step to remove the specificity of a caption that is relevant only in the retrieved image such as the name of a city. The model reported by Socher et al. [37] uses dependent representations and neural networks to embed images and sentences together, into a common vector form. This approach shows how to map sentence representations from recursive networks into the same space as images. Vinyals et al. [38] demonstrated the effectiveness of storing contextual information in a recurrent layer, and developed a generative approach based on a combination of Convolutional and Recurrent Neural Networks, to generate image captions monitoring their output on the image features extracted by a convolutional neural network. This approach uses the MS COCO dataset that contains 5000 images with 40 reference sentences to enhance the accuracy of automatic measures. Kiros et al. [39] used two separate pathways (for images and text) to define a joint embedding, even if they can generate text. They proposed a different architecture using the hidden state of an LSTM (Long Short-Term Memory) encoder at time T as the encoded representation of the length T input sequence. It maps this sequence representation and combines it with the visual representation of a modern visual Convnet model, a joint space is obtained with a separate decoder predicts words.

Mao et al. [40] opened new prospects for bidirectional methods that retrieve images based on a textual input, or sentences from a given image. They developed powerful methods of jointly learning from image and text inputs to form

higher-level representations from models such as convolutional neural networks (CNNs). They tested their methods on object recognition and word embedding taken from a large-scale text corpus. They proposed a system using Convolutional Neural Networks to extract image features, and Recurrent Neural Networks for sentences, with an interaction performed in a multimodal common layer.

Figure 4 shows a precise segmentation of two scenes extracted from PETS challenge. The scene contains static elements that do not change over time (i.e., buildings, grass, electric poles, roads, trees, car parks, restrictive roads). All these concepts belong to our ontology.

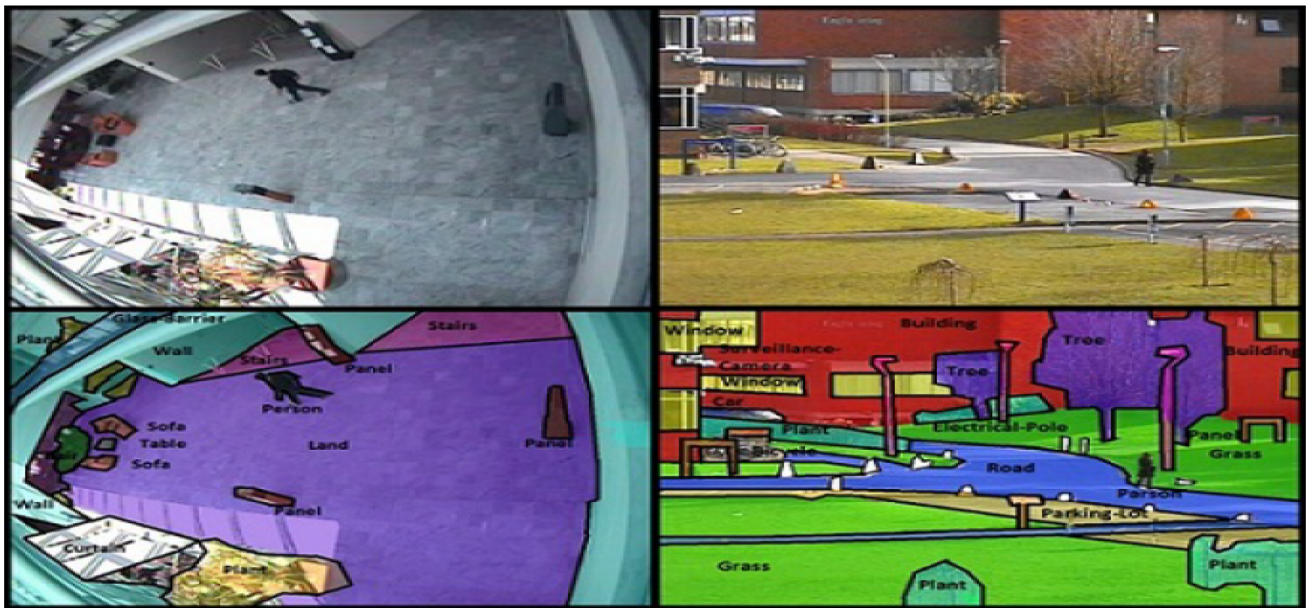


Fig. 4 Scene description from PETS 2004 and PETS 2012 challenges

5.3 Video description

Like image description, the video description process represents operations that describe objects acting in a video clip. However, unlike the images that are static, the videos require the information dealing with dynamic and temporal changes of the structure along with translation into a natural language description.

Many works have led to greater exploration of video description applications. Yao et al. [41] developed detailed models using Long Short-Term Memory (LSTM) description capable to select the most relevant temporal segments in a video and to incorporate 3D CNN by generating sentences. In their work, two types of encoders are tested: one is a simple frame-wise application of the pre-trained convolutional network, while the other is a 3D convolutional network. Rohrbach et al. [42,43] used an approach based on statistical machine translation that produces descriptions of videos containing several people cooking in the same kitchen, with the possibility to go from an intermediate semantic representation to sentence generation. Sentences are generated starting from a semantic role representation of high-level concepts such as the actor, action and object. Venugopalan et al. [44] applied the neural approach to static image caption generation, and used an LSTM decoder type for automatic video description generation tasks. They used a convolutional neural network to extract appearance features from each frame of an input video clip. Due to its complete and coherent representation, our proposed ontology could be used easily in all video description works.

5.4 Video event indexing

Video indexing offers advanced computer vision capabilities to efficiently and automatically categorize and search events in large datasets. It describes the process of events detection in the video surveillance domain. The consistent and the diversity of our ontology in terms of video surveillance concept representation can incorporate video event indexing and retrieval systems. In Sect. 7, we present an application of our ontology to video surveillance event indexing, using the PETS 2012 and TRECVID 2016 datasets We have selected five event recognition tasks to depicts the efficiency of the proposed OVIS system.

6 OVIS system architecture

The ontology approach is the effective way to support the event indexing process in the video surveillance domain. It represents the core module in the global architecture of the OVIS system as shown in Fig. 5. Its main purpose is to ensure the video sequence indexing process from the first step using the blobs extraction module for extracting blobs bounding box features to the last step of events identification and video sequences indexation.

The indexing process of video sequences shown in Fig. 5, starts when the video analysis module extracts the different blobs bounding boxes from the video sequence using a background subtraction method with some low-level properties such as Top Left Point X, Top Left Point Y, Width, and Length. The ontology considers these bounding box features

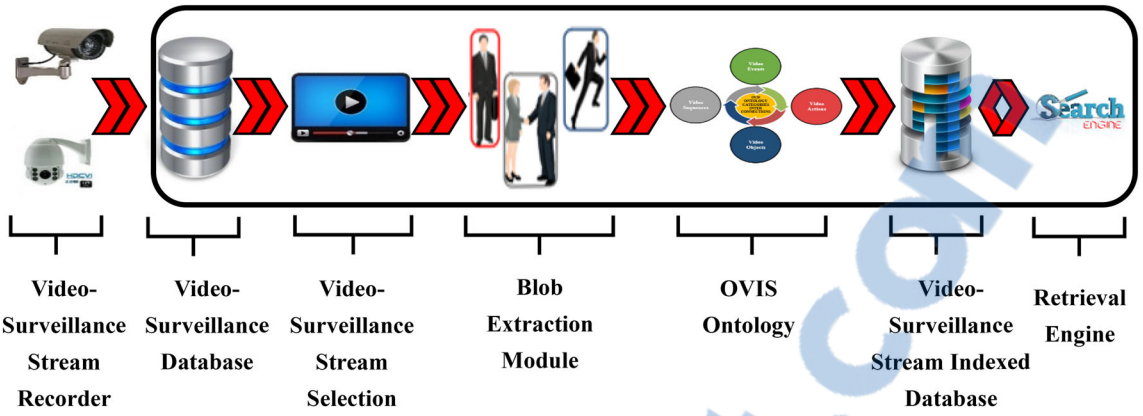


Fig. 5 OVIS system Architecture

as an input, and organize them to create DataProperty and ObjectProperty, frame, video sequence, etc. Then, the reasoner of our ontology gives them the correct order using a set of SWRL rules, index this video sequence into the appropriate video event class according to the behavior of its objects with start and end event frames. Finally, the video surveillance stream will be indexed and stored in the database for future needs. For retrieval purpose, the OVIS system allows the search of all indexed videos in the video surveillance stream indexed database using key words expressed as events names (e.g., walking, running, splitting, formation and local dispersion). For example, if we like to retrieve walking events, we will use walking as keyword and the OVIS system will return all sequences indexed with a walking event.

6.1 Blobs extraction module

Blobs represent features that usually have a large coverage area and have proved to be better than points, corners or edges, due to the full occlusion of the subject. Several algorithms could be used to collect the blobs. The background subtraction algorithm will classify the pixels of the input image into foreground and background. The blobs are extracted by collecting the foreground pixels that belong to a single connected component. Optical flow can be used by extracting the characteristics of each pixel in each motion image. These flows are then grouped into blobs with coherent motion and are modeled by a mixture of multivariate Gaussians. The optical flows are useful to characterize each moving pixel according to certain features of the flows' vectors. In the present work, a background subtraction method is used to extract blobs that occur in each frame with their bounding boxes. These features represent the input of our SWRL approach for event detection purposes.

For video analysis, we have used the OpenCV [45] library to extract low-level features of blobs bounding boxes such as top left coordinates, height and width. Therefore, all these features will be used as an input of the OVIS system.

6.2 Methodology for populating the ontology

Our solution is open-source based on the Pellet reasoner and Protege2000 (5.0.0 version) application. However, the biggest challenge was to fill the population of the ontology. We have performed a reverse engineering of Protege2000 to understand how to create the automated filling of individuals with Data_Property and Object_Property features. In each OWL document generated by Protege2000, we found different properties like individuals, DataProperty and ObjectProperty. Therefore, the solution represents the creation of the parser that reads and extracts different data from the output file of the image/video processing. Then, the parser opens and includes the right tag of the OWL file based on our ontology modeling, the different individuals represented as bounding boxes generated with their Data_Properties and Object_Properties. Finally, the population of the ontology (individuals represented as bounding boxes generated from image/video processing) was already filled with their "Data_Property and Object_Property" opened with the new generated OWL file.

6.3 SWRL rules

To test the efficiency of the proposed approach, different events are addressed together with more than 300 SWRL rules (see web link[1] for some example of SWRL rules), such as:

- Group running and walking events: In each image, the motion magnitude identifies the difference between these

[1] http://ovis-system-information.000webhostapp.com/.

two events. For instance, a high-magnitude event means running, while a low-magnitude event means walking. The detection is performed by defining an experimental threshold or using a classifier with a feature such as the average speed of movement. In our case, we used an experimental threshold.

- Group formation and splitting events: The position, orientation and speed of the groups are the main factors determining the accuracy of the events.
- Group Local dispersion events: the positions and the evolving size of the group over frames determine the accuracy of the events.

We used the rule plugin of Protege [46] to write the inference rules of our engine in the SWRL language, and the Pellet reasoner [47] to infer all the events. These SWRL rules are divided into three categories: distance, tracking and event rules.

6.3.1 SWRL distance rules

They consist of generating all major bounding box in each frame of the video sequence. These rules check distances between detected bounding boxes in the current frame. Neighboring Bounding boxes are grouped into a major one.

Figure 6 depicts an example of a situation for grouping two bounding boxes detected with the blobs extraction module into a major one. An SWRL distance rule verifies whether these bounding boxes could be grouped into the same major bounding box or not. The Pellet Reasoner takes the decision of inferring or not across the right side and checks the left side of the SWRL rule (before the arrow).

6.3.2 SWRL tracking rules

These rules consist of generating all the different Group_Of_Person instances using the results of major bounding boxes created by the previous rules (SWRL Distance rules) to detect the start/end position of each group and other parameters in the video sequence.

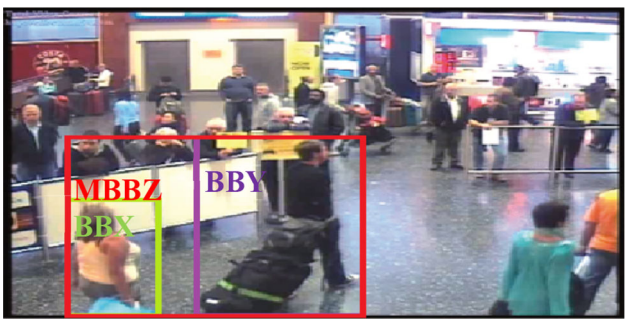


Fig. 6 An illustration of a situation for grouping two bounding boxes into a major one (TRECVID 2016)

Figure 7 illustrates an example of a situation describing the tracking of the identified group GPY between two successive frames (FZ and FZ+1).

6.3.3 SWRL event rules

This category of rules is used to detect the appropriate events. Therefore, the behavior of the group identified in the previous category (SWRL Tracking rules) is analyzed. An example of a splitting event is illustrated in Fig. 8 below where an event SWRL rule is used for verifying whether the identified group GPZ splits or not into two groups (GPX and GPY) between two successive frames (FZ and FZ+1).

After presenting the three categories of our SWRL rules, the next paragraph demonstrates all the strategies used for reasoning and inferring the different events.

6.3.4 Reasoning strategies

In the aim of inferring the different events presented above, we create different SWRL rules with a strategy that follows all these 16 steps in the reasoning process:

1. Inferring the four coin points of each bounding box (x, y) detected with the low-level extractor module.
2. Inferring the major bounding box in the case of frame containing only one bounding box.
3. Inferring the four coin points of each major bounding box (x, y) (case of frame containing only one bounding box).
4. Inferring an ID for each bounding box detected with respect to its position in the frame (the one who is most to the right, will have the ID number one, then the one who comes to his left will have the ID number two and so on).
5. Inferring the majors bounding boxes in the case of frame containing two bounding boxes.
6. Inferring the four coin points of each major bounding box (x, y) (case of frame containing two bounding boxes).
7. Inferring the majors bounding box in the case of frame containing three bounding boxes or more:

   7.1. Comparison between Bounding Boxes having as ID number one and two to extract all the blocks with their MIN and MAX (x, y).
   7.2. Comparison of the generated blocks with different Bounding Boxes of frame and inferring an FID (Final ID) for each Bounding Box.
   7.3. Comparison between Bounding Boxes with final ID number one and final ID number two and extract

Fig. 7 An illustration of checking if the MBB detected in frame FZ+1 represents the same GPY in frame FZ (PETS 2012)

Fig. 8 An illustration of a situation for checking if the group GPZ is split or not (PETS 2012)

either an MBB (Major Bounding Box) in the case of a large distance or a TBB (Temporary Bounding Box) in the case of a small distance.

7.4. Comparison between a TBB and the rest of Bounding Boxes with respect to the order of the final ID and extract either an MBB (Major Bounding Box) or a new TBB (Temporary Bounding Box) with the same strategy of distance noted in the last point.

7.5. Inferring the four coin points of each major bounding box (x, y) generated above.

8. Measuring the centroid of each Major bounding box.
9. Search the first frame that contains a Major Bounding Box and inferring a TGP (Temporary Group-of-Person).
10. Inferring groups with their properties: Started Frame, Ended frame, in the normal case.
11. Inferring Groups with their properties: Started Frame, Ended frame, in the case of final frame.
12. Inferring Groups with their properties: Started Frame, Ended frame, in the case of empty frame.
13. Search the first frame that contains a Major Bounding Box after browsing an empty frame.
14. Inferring relations between detected groups.
15. Checking if it is not a false detection.
16. Inferring the type of event.

7 Results and discussions

To test the efficiency of the OVIS system inspired by TRECVID and PETS, we selected five events (walking, running, split, formation and local dispersion). We developed an application in the Java environment that handles all the steps of our indexing and retrieval system. The process started with the selected video and ended with the indexing results. The tests were performed on a machine with Intel Core I7 CPU and 16 GB RAM, under Windows 8. We considered three types of evaluations to check the performance of the OVIS system:

7.1 Evaluation based on the events

The first type of evaluation was based on the number of events returned by the OVIS system; it is carried out by many metrics, such as Precision, Recall, F-measure, FP (False Pos-

itive), FN (False Negative), TP (True Positive) and TN (True Negative). We considered these measures as follows:

- Precision = Number of detected videos that contain the event/Number of videos indexed with the event.
- Recall = Number of detected videos that contain the event/Number of all videos in Database that contain the event.
- F-measure = 2 * ((Precision * Recall)/(Precision + Recall)).
- TP = Number of videos indexed with the event and containing it.
- TN = Number of videos not indexed with the event and not containing it.
- FP = Number of videos indexed with the event and not containing it.
- FN = Number of videos not indexed with the event and containing it.

We selected two challenges (PETS 2012 and TRECVID 2016) to evaluate our event detection system.

In the first challenge, we used 16 videos (with 4240 frames) that were used to describe different events and we carry out all the metrics noted above.

In the second one, an 11-h subset of the multi-camera airport surveillance domain evaluation data is used with the following evaluation metrics: select Precision, Recall and F-measure.

7.1.1 PETS 2012 challenge

Table 7 shows the indexing results for each event. We consider 13 walking, 8 running, 4 splitting, 4 formation and 4 local dispersion events as the ground truth videos, given by the first column. The second column displays the number of videos that the OVIS system indexes in each event; the final column shows the number of videos containing the effective events among those indexed by OVIS.

Discussion 1 In Table 8, we summarized the statistics of the obtained data from the full dataset. On one hand, the events walking and local dispersion provide excellent results and reach 100% of precision. Therefore, these results mean that the number of detected videos by OVIS system that contains walking and local dispersion events is equal to the number of videos indexed with these events. On the other hand, the events of running, splitting, formation exceed 50% of precision. So, we can conclude that the number of detected videos by OVIS system that contains running, splitting, and formation events is equal to at least the half of the number of videos indexed with these events. As illustrated, the event local dispersion reaches also an excellent result of 100% in recall. This means that the OVIS system does not miss any local dispersion event. Moreover, the recall of event walking, running, splitting and formation provides good result and is at least above 62%. Following the result of precision and recall, the F-measure metric expresses the relation precision/recall. Consequently, the F-measure metric provides excellent result in walking and local dispersion events and a good one in the rest of events. The metrics (FP, FN, TP and TN) give the accuracy of the indexing process generated by of the OVIS system. Thus, this accuracy is considered as excellent when the number of FP and FN is very low and the number of TP and TN is very high. As described in Table 8, these metrics provide good results in general. For example, walking event generates good results with 3 videos not indexed with

Table 7 OVIS indexing results of the five different events

Results events	Number-of-all-video-in-Database-that-really-contain-the-event (ground truth)	Number-of-video-returned	Number-of-video-returned-that-really-contain-the-event
Walking	13	11	11
Running	08	09	05
Split	04	06	03
Formation	04	06	03
Local dispersion	04	04	04

Table 8 The obtained results for the five different events

Measures events	Precision (%)	Recall (%)	F-measure(%)	FP	FN	TP	TN
Walking	100	84	91	00	02	11	03
Running	55	62	58	04	03	05	04
Splitting	50	75	60	03	01	03	09
Formation	50	75	60	03	01	03	09
Local dispersion	100	100	100	00	00	04	12

Table 9 Comparison of different detection event approaches, NC (Not communicated), ND (Event not detected)

Event	Metrics	The OVIS system	Utasi et al. [28]	Chan et al. [29]
Group walking	Precision	1	ND	0.87
	Recall	0.84	ND	NC
Group running	Precision	0.55	0.99	0.75
	Recall	0.62	0.99	NC
Group splitting	Precision	0.5	0.6	0.74
	Recall	0.75	1	NC
Group formation	Precision	0.5	ND	0.6
	Recall	0.75	ND	NC
Group local dispersion	Precision	1	ND	0.8
	Recall	1	ND	NC

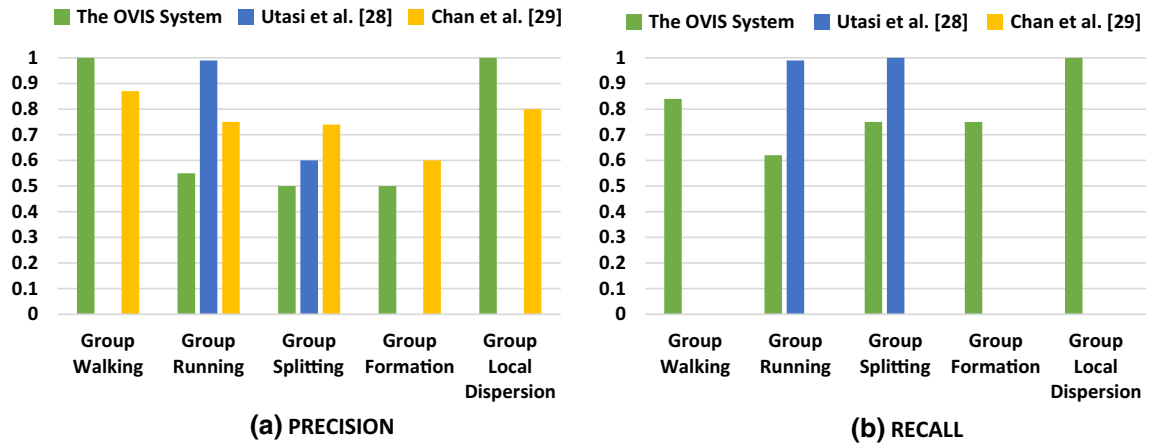


Fig. 9 Comparison of different detection event approaches under graphical forms, **a** precision, **b** recall

the event and not detected among 3 videos (TN), 11 videos indexed with the event and containing it among 13 videos (TP), 2 videos not indexed with the event and containing it (FN) and no video not indexed with the event and containing it (FP). This is an evidence that the OVIS system using SWRL rules with their different categories (Distance rules, Tracking rules, Event rules) successively, opens new prospects without using traditionally methods (SVM, KNN, etc.). The use of this complete video surveillance ontology is efficient in this indexing process that represents one among other domain applications addressed by this ontology.

Comparison with other approaches For the aim of evaluating the OVIS system based on SWRL rules, we compared it with two other studies [28,29] that use the same video sequences from the PETS 2012 challenge dataset and handle the same event detection purpose: Group

Walking, Group Running, Group Splitting, Group Formation and Group Local Dispersion.

Table 9 compares the results obtained with the OVIS system and those reported in [28,29], using the Dataset PETS 2012. Figure 9 illustrates the Precision/Recall of our approach compared with the others.

Our approach based on SWRL rules detects all events, while Utasi et al. [28] method detects only two events (Group Running and Group Splitting). They did not try to identify the missing events although they claim that their approach was able to detect the other events without providing any indications. The method of Chan et al. [29] cannot include two events simultaneously while OVIS offers the possibility of processing and detecting two or more events at the same time, as illustrated in Fig. 11 below. Furthermore, the approach of Chan et al. [29] does not provide some important results as indicated by NC in Table 9. Even if the precision of [29] is better than OVIS in few cases, they did not present their recall.

These observations lead us to conclude that the negative points noted above (not detecting all events, not processing two events simultaneously and not providing recall results) prove that our approach based on SWRL rules and created from the presented ontology is strong. We could also add new events just by creating the associated SWRL rules.

Table 10 OVIS indexing results of the three different events

Results events	Number-of- event (ground truth)	Number-of-all-event-returned-by-OVIS	Number-of-correct-event-returned-by-OVIS
People meet	323	411	303
People split up	176	203	173
Person runs	63	97	58

Table 11 Comparison of different detection event approaches

Event	Metrics	The OVIS system	Markatopoulou et al. [30]	Zhao et al. [31]
People meet	Precision	0.74	0.02	0.34
	Recall	0.94	0.92	0.18
	F-measure	0.83	0.04	0.23
People split up	Precision	0.86	0.01	0.32
	Recall	0.98	0.98	0.2
	F-measure	0.92	0.02	0.25
Person runs	Precision	0.6	0.01	0.67
	Recall	0.92	0.97	0.35
	F-measure	0.73	0.02	0.46

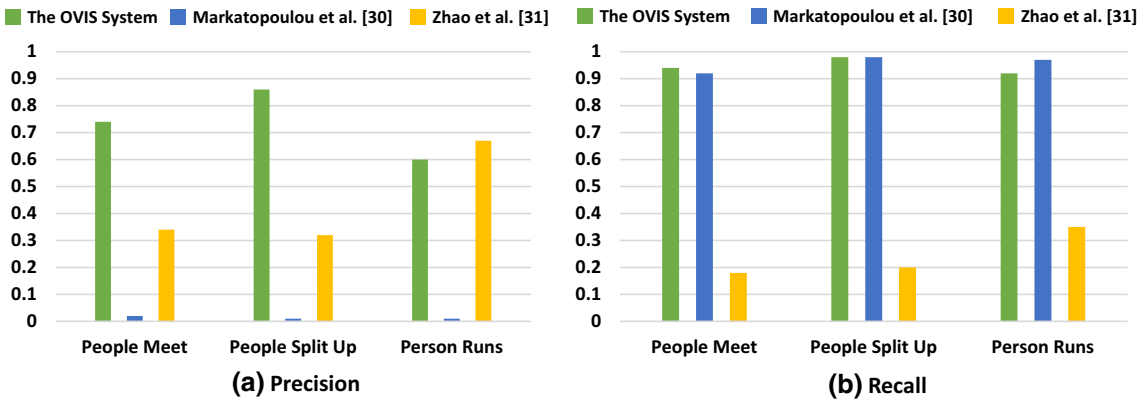


Fig. 10 Comparison of different detection event approaches under graphical forms, **a** precision, **b** recall

7.1.2 TRECVID 2016 SED task

For the aim of evaluating our OVIS system in various contexts, we perform a comparison with two other approaches [30,31] that participate in the SED task of TRECVID 2016 challenge. The SED task regroups seven types of events: PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing. In our case studies, our OVIS system with it different SWRL rules handles only three of them: PersonRuns (Running), PeopleMeet (Formation), PeopleSplitUp (Splitting). In this way, we compare results obtained from our OVIS system with those presented in [30,31] representing these three events.

Table 10 shows the indexing results of PeopleMeet, PeopleSplitUp and PersonRuns events returned by our OVIS system. In our case, the OVIS system based on SWRL rules inferring method return the events of Formation, Splitting and Running and we consider them as PeopleMeet, PeopleSplitUp and PersonRuns, respectively. We consider 323 PeopleMeet, 176 PeopleSplitUp and 63 PersonRuns events as the ground-

truth videos, given by the first column. Each event was detected with its start and end frames that provide the correct detection or false alarm event.

Table 11 illustrates the results obtained with the OVIS system and those reported in [30,31], using the Dataset TRECVID 2016 SED task, where Fig. 10 shows the Precision/Recall metrics of our approach compared with the others.

First, our approach based on SWRL rules has a better ratio Precision/Recall (F-measure) compared with those exposed in [30,31]. This point means that our system detects the

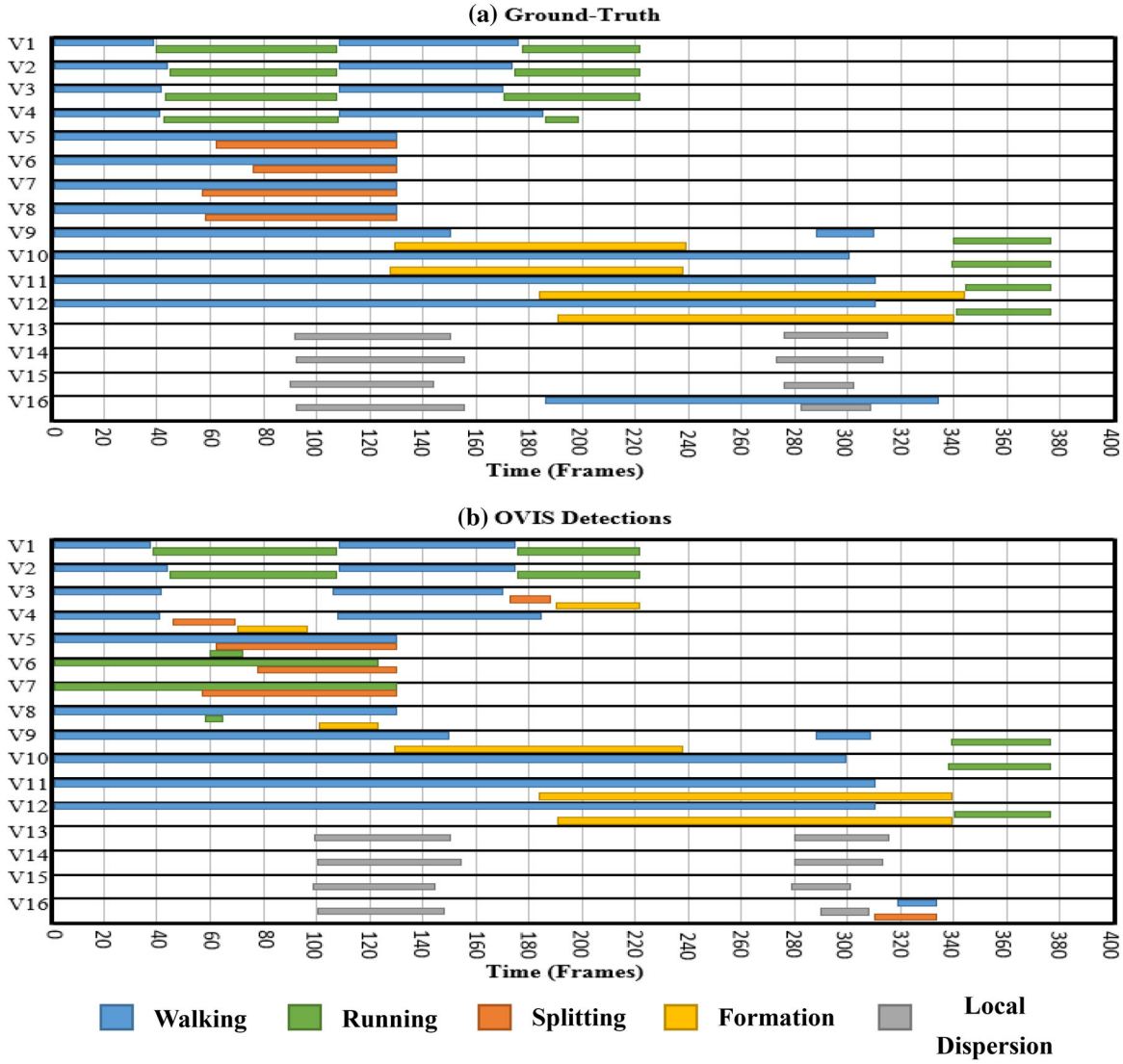


Fig. 11 Frame timing events of sixteen videos (Ground Truth and results returned by the OVIS system)

majority of events related to ground truth without detecting a great number of false alarms. Secondly, the approach Markatopoulou et al. [30] detects the majority of correct event like our OVIS system but at the same time a great number of false alarms. We can conclude that this approach does not miss a large number of correct events but their precision is very low compared with OVIS system. Third, the method of Zhao et al. [31] expresses acceptable results in precision metrics but at the same time misses a great number of correct detection events compared with OVIS system. These observations lead us to conclude that all weaknesses detected in [30,31] (miss a great number of events, detect a great number of false alarms) prove that our approach based on SWRL rules and created from this ontology is strong. We could also add new events like CellToEar, ObjectPut Embrace, and Pointing or other ones by creating the associated SWRL rules.

7.2 Evaluation based on the frame timing (PETS 2012)

The second type of evaluation is based on frame timing using an overlay (in frame number) of 10 frames to evaluate the performance of our system. To meet this end, three kinds of situations are considered:

- Too Early: Our system detects events before they really start (ground truth) with 10 frames.
- On Time: Our system detects events exactly when they start.
- Too Late: Our system detects the events after they really start with 10 frames.

Figure 11 represents the frame timing events of sixteen videos (Ground Truth and results returned by the OVIS sys-

Table 12 Illustration of the inferring results representing Allen's relations provided by the OVIS system

Video_Sequences	Allen's relations detected by OVIS
Video 1	Took_Place_Before_R/Took_Place_After_R/Met_R/Met_By_R
Video 2	Took_Place_Before_R/Took_Place_After_R/Met_R/Met_By_R
Video 3	Took_Place_Before_R/Took_Place_After_R/Met_R/Met_By_R
Video 4	Took_Place_Before_R/Took_Place_After_R/Met_R/Met_By_R
Video 5	Finished_R/Finished_By_R/Overlapped_R/Overlapped_By_R/ Took_Place_During_R/Contained_R
Video 6	Overlapped_R/Overlapped_By_R
Video 7	Finished_R/Finished_By_R
Video 8	Took_Place_During_R/Contained_R/Took_Place_Before_R/ Took_Place_After_R
Video 9	Overlapped_R/Overlapped_By_R/Took_Place_Before_R/ Took_Place_After_R
Video 10	Took_Place_Before_R/Took_Place_After_R
Video 11	Overlapped_R/Overlapped_By_R
Video 12	Overlapped_R/Overlapped_By_R/Met_R/Met_By_R
Videos 13 /14/ 15	
Video 16	Took_Place_Before_R/Took_Place_After_R/Met_R/Met_By_R

tem). The annotation for ground truth is done manually by watching the entire video, and represents in each case the events that occur with their start/end frames.

Discussion 2 Figure 11 above illustrates the results for the frame timing events obtained from the output of the OVIS system compared with the ground truth. The first advantage is the detection of at least one correct event in each video sequence analyzed by our system. Moreover, the second positive point is that all the correct events are detected on time without exceeding the overlay of 10 frames. These positive points demonstrate that our approach (based on the three types of SWRL rules from blobs features until event detection) works correctly at least once, and detects the right event at the right moment, where [28–31] fail in this task. However, the weakness of the OVIS system is the occasional events confusions that generate incorrect events detections. This is generally due to these two examples cases:

1. The emergence of objects (not relevant) detected in the video sequence by the low-level feature analysis module; their behavior leads our system to detect a wrong event. For example, when an image/ video processing detects a moving tree shadow and considers it as bounding box of pertinent blobs. This wrong detection leads OVIS system to detect a wrong event in general.
2. The incorrect splitting and formation of detected objects caused only by their walking or running speed. For example, when in the same Group-Of-Person, we detect different human speeds and in presence of walking or running events, OVIS can detect a wrong Splitting or Formation event.

We are confident that these weaknesses can be solved in a future work by adding new SWRL rules and improving initial blob detection and characterization. We not only expect to improve the obtained results, but also predict and extend the tests for detecting new events.

7.3 Evaluation based on Allen's interval algebra (PETS 2012)

The third type of evaluation is based on Allen's interval algebra 3 [48] for modeling many possible situations such as "X takes place before Y", "X overlaps Y", "X meets Y", …. (Where X and Y represents two different events detected in the same video sequences). For this purpose, another kind of SWRL rules were developed, and aim to infer the different relations between events detected in the same video sequence. The results of these SWRL rule were saved as a Boolean Data_Property related to Video_Sequences individuals (For example of relation: Took_Place_Before_R). Therefore, all the relations expressed as Data_Property inferred with these SWRL rules take true as value. For inferring the different relations between events detected in the same Video_Sequences, the start frame and end frame of each event represents the key aspects. Furthermore, the detection of five frames between two different successive events was considered in our case as "Met_R/Met_By_R" relations, and therefore up to five frames as "Took_Place_Before_R/Took_Place_After_R" relations. However, when the Video_Sequences contain only one event detected in several times, like Videos (13, 14, and 15) in Table 12, no relations was inferred.

Therefore, we expect seven possible relations with their opposite situations:

- X took place before Y: Where event X finished before when event Y started.
- Y took place after X: Where event Y started after when event X finished.
- X met Y: Where event X finished at the same time when event Y started.
- Y met by X: Where event Y started at the same time when event X finished.
- X overlapped Y: Where event X started before event Y and event Y finished after event X.
- Y overlapped by X: Where event Y started after event X and event X finished before event Y.
- X started Y: Where event X started at the same time with event Y and finished before event Y.
- Y started by X: Where event Y started at the same time with event X and finished after event X.
- X took place during Y: Where event X started after event Y and finished before event Y.
- Y contained X: Where event Y started before event X and finished after event X.
- X finished Y: Where event X started after event Y and finished at the same time with event Y.
- Y finished by X: Where event Y started before event X and finished at the same time with event X.
- X equaled to Y: Where event X started and finished at the same time with event Y.

Discussion 3 Table 12 above illustrates all the Allen's relations obtained from the output of the OVIS system. The first advantage is the detection of all the correct relations in the sixteen video. For example, in Video 12 "V12" illustrated in Fig. 11 above (OVIS Detections part) handles three kinds of events (Walking, Formation, and Running). The different relations that we can observe is that:

- First, event Walking started before event Formation and event Formation finished after event Walking (X overlapped Y) and (Y overlapped by X).
- Second, event Formation finished at the same time when event Running started (X met Y) and (Y met by X).

However, all these Allen's relations are inferred by the OVIS system like described in Table 12 (Video 12). The second advantage is no detection of any relation in presence of only one event detected in a video sequence, is done. For example, video 13, video 14 and video 15 illustrated in Fig. 11 above, we can find only Local-Dispersion event. Consequently, any relation was inferred by OVIS system as presented in Table 12. These positive points demonstrate that our approach based of SWRL is able to handle correctly Allen's relations whereas this point is not generated or presented also in [28–31].

8 Conclusions

Nowadays, video surveillance systems are part of our daily life, because of their role to ensure security and safety (i.e., allowing human behavior to be studied among the population). Thus, many research works have tried to develop an efficient system to index a very large volume of data accurately. In the present work, we have proposed a complete and coherent video surveillance ontology and a rule-based approach to detect multiple object events or crowd events (e.g., Group walking, Group splitting, etc.). In fact, we have described the link between the four main categories composing our ontology (Video_Sequences, Video_Objects, Video_Events, Video_Actions), that are in interaction.

Our video surveillance ontology covers a very large number of objects and events happening in the video surveillance domain, as well as exhibiting a large dimension taking into consideration new concepts that represent events in the industrial domain.

Furthermore, we implemented the OVIS indexing and retrieval system, based on a complete video surveillance ontology and SWRL rules in the middle and high level of indexing process. Moreover, we tested OVIS with videos selected from the PETS 2012 and TRECVID 2016 Challenges. In this way, we obtained very promising results.

Moreover, the strengths of our approach are as follows:

- The competitive level results obtained with the different types of evaluations such as: evaluation based on the events, evaluation based on the frame timing, evaluation based on Allen's interval algebra.
- The facility of creating and using SWRL rules or adding new ones when new events occurs. This point allows the research community to address many future prospects in the domain ontology-based video surveillance indexing and retrieval systems.

The weaknesses of our approach stem from the requirement of manual reproduction of SWRL rules when new events occur and the lack of the uncertainty management by the OVIS system in image/video processing.

In our future work, we will extend the OVIS system by considering other events that could occur in the video surveillance domain. This will be possible by adding new SWRL rules and testing them using other datasets. In addition, we plan to use the neuronal network formalism to reproduce others SWRL rules, and use the Shannon's normalized entropy function for modeling the uncertainty associated to image/video processing.

References

1. Kless D, Jansen L, Lindenthal J, Wiebensohn J (2012) A method for reengineering a thesaurus into an ontology. In: Frontiers in artificial intelligence and applications (FAIA), pp 133–146
2. Badii A, Lallah C, Zhu M, Crouch M (2009) The dream framework: Using a network of scalable ontologies for intelligent indexing and retrieval of visual content. In: International conference on web intelligence and intelligent agent technology (WI-IAT), pp 551–554
3. Rodrguez-Muro M, Calvanese D (2012) High performance query answering over DL-Lite ontologies. In: International conference on principles of knowledge representation and reasoning (KR), pp 308–318
4. Scherp A, Saathoff C, Franz T, Staab S (2011) Designing core ontologies. J Appl Ontol 03:177–221
5. Benmokhtar R, Huet B (2011) An ontology-based evidential framework for video indexing using high-level multimodal fusion. Multimed Tools Appl 55(3):1–27
6. Rector A, Brandt S, Drummond N, Horridge M, Pulestin C, Stevens R (2012) Engineering use cases for modular development of ontologies in owl. J Appl Ontol 02:113–132
7. Smith B, Ceusters W (2010) Ontological realism as a methodology for coordinated evolution of scientific ontologies. J Appl Ontol 03(4):139–188
8. Hernandez-Leal P, Escalante HJ, Sucar LE (2017) Towards a generic ontology for video surveillance. In: Applications for future internet
9. Kara S, Alan Z, Sabuncu O, Akpnar S, Cicekli NK, Alpaslan FN (2012) An ontology-based retrieval system using semantic indexing. Inf Syst J 04:294–305
10. Mossakowski T, Lange C, Kutz O (2013) Three semantics for the core of the distributed ontology language. In: International joint conferences on artificial intelligence (IJCAI), pp 3027–3031
11. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. Multimed Tools Appl 02:313–337
12. Bagdanov AD, Bertini M, Del Bimbo A, Serra G, Torniai C (2007) Semantic annotation and retrieval of video events using multimedia ontologies. In: International conference on semantic computing (ICSC), pp 713–720
13. Bertini M, Del Bimbo A, Torniai C, Grana C, Cucchiara R (2007) Dynamic pictorial ontologies for video digital libraries annotation. In: 1st ACM workshop on the many faces of multimedia semantics, pp 47–56
14. Bertini M, Del Bimbo A, Serra G (2008) Learning ontology rules for semantic video annotation. In: 2nd ACM workshop on multimedia semantics, pp 1–8
15. OConnor M, Knublauch H, Tu S, Grosof B, Dean M, Grosso W, Musen M (2005) Supporting rule system interoperability on the semantic web with SWRL. In: 4th international semantic web conference (ISWC), pp 974–986
16. Xue M, Zheng S, Zhang C (2012) Ontology-based surveillance video archive and retrieval system. In: 5th International conference on advanced computational intelligence (ICACI), pp 84–89
17. Lee J, Abualkibash MH, Ramalingam PK (2008) Ontology based shot indexing for video surveillance system. In: Innovations and advanced techniques in systems, computing sciences and software engineering, pp 237–242
18. Snidaro L, Belluz M, Foresti GL (2007) Representing and recognizing complex events in surveillance applications. In: IEEE international conference on advanced video and signal-based surveillance (AVSS), pp 493–498
19. Calavia L, Baladrn C, Aguiar JM, Carro B, Sanchez-Esguevillas A (2012) A semantic autonomous video surveillance system for dense camera networks in smart cities. Sensors 12:10407–10429
20. Papadopoulos GT, Mezaris V, Kompatsiaris I, Strintzis MG (2007) Ontology-driven semantic video analysis using visual information objects. In: International conference on semantic and digital media technologies, pp 56–69
21. Saad S, Beul DD, Said M, Pierre M (2012) An ontology for video human movement representation based on benesh notation. In: IEEE international conference on multimedia computing and systems (ICMCS), pp 77–82
22. Trochidis I, Tambouris E, Tarabanis K (2007) An ontology for modeling life-events. In: IEEE international conference on services computing (SCC), pp 19–20
23. Bohlken W, Neumann B (2009) Generation of rules from ontologies for high-level scene interpretation. In: Lecture notes in computer science, pp 93–107
24. Nevatia R, Hobbs J, Bolles B (2004) An ontology for video event representation. In: Computer vision and pattern recognition (CVPR), pp 119–128
25. Francois ARJ, Nevatia R, Hobbs J, Bolles RC, Smith JR (2005) VERL: an ontology framework for representing and annotating video events. IEEE Multimed 12:76–86
26. Bai L, Lao S, Zhang W, Jones GJF, Smeaton AF (2008) Video semantic content analysis framework based on ontology combined mpeg-7. In: Lecture notes in computer science, pp 237–250
27. SanMiguel JC, Martinez JM, Garcia A (2009) An ontology for event detection and its application in surveillance video. In: IEEE international conference on advanced video and signal-based surveillance (AVSS), pp 220–225
28. Utasi A, Kiss A, Sziranyi T (2009) Statistical filters for crowd image analysis. In: Performance evaluation of tracking and surveillance workshop, at CVPR, pp 95–100
29. Chan AB, Morrowand M, Vasconcelos N (2009) Analysis of crowded scenes using holistic properties. In: 11th IEEE international workshop on performance evaluation of tracking and surveillance (PETS)
30. Zhao Z, Wang M, Xiang R, Zhao S, Zhou K, liu M, He S, Zhu Y, Zhao Y, Su F (2016) BUPT-MCPRL, at TRECVID
31. Markatopoulou F, Moumtzidou A, Galanopoulos D, Mironidis T, Kaltsa V, Ioannidou A, Symeonidis S, Avgerinakis K, Andreadis S, Gialampoukidis I, Vrochidis S, Briassouli A, Mezaris V, Kompatsiaris I, Patras I (2016) ITI-CERTH, at TRECVID
32. Kazi Tani MY, Ghomari A, Belhadef H, Lablack A, Bilasco IM (2014) An ontology based approach for inferring multiple object events in surveillance domain. In: IEEE science and information conference (SAI), pp 404–409
33. Kazi Tani MY, Ghomari A, Lablack A, Bilasco IM (2015) Events detection using a video-surveillance ontology and a rule-based approach. In Computer vision + ONTology applied cross-disciplinary technologies workshop (CONTACT) in conjunction with European conference in computer vision (ECCV), pp 299–308
34. PETS. Pets 2012 challenge. http://www.cvg.reading.ac.uk/PETS2012/a.html
35. TRECVID. TRECVID 2016 challenge. http://www-nlpir.nist.gov/projects/tv2016/tv2016.html
36. Kuznetsova P, Ordonez V, Berg T, Choi Y (2014) Treetalk: composition and compression of trees for image descriptions. In: Transactions of the association for computational linguistics (TACL), pp 351–362
37. Socher R, Karpathy A, Le VQ, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist 2:207–218

38. Vinyals O, Toshev A, Bengio S, Erhan D (2014) Show and tell: a neural image caption generator. arXiv:1411.4555
39. Kiros R, Salakhutdinov R, Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539
40. Mao J, Xu W, Yang Y, Wang J, Yuille AL (2014) Explain images with multimodal recurrent neural networks,.arXiv:1410.1090
41. Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. In: IEEE international conference on computer vision (ICCV)
42. Rohrbach A, Rohrbach M, Qiu W, Friedrich A, Pinkal M, Schiele B (2014) Coherent multi-sentence video description with variable level of detail. In: German conference on pattern recognition (GCPR)
43. Rohrbach M, Qiu W, Titov I, Stefan T, Pinkal M, Schiele B (2013) Translating video content to natural language descriptions. In: IEEE international conference on computer vision (ICCV)
44. Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney RJ, Saenko K (2014) Translating videos to natural language using deep recurrent neural networks. arXiv:1412.4729
45. OpenCV. The OpenCV API. http://docs.opencv.org/3.3.0/
46. Protege. The protege project. http://protege.stanford.edu
47. Sirin EB, Parsia B, Cuenca Grau B, Kalyanpur A, Katz Y (2003) Pellet: a practical OWL-DL reasoner. J Web Semantics 5:51–53
48. Allen JF (1983) Maintaining knowledge about temporal intervals. Commun ACM 26:832–843

Résumé

De nos jours, le nombre de vidéos de différents types et domaines est devenu impressionnant grâce à l'évolution technologique où on peut citer à titre d'exemples les vidéos des journaux télévisés, du Web, du sport, ou dans le domaine des divertissements comme des films, des séries télévisées, des documentaires ou encore les vidéos de la surveillance. Cette diversité et grande masse de vidéos disponibles sur Internet nécessite de développer des outils qui permettent d'organiser ce type de données afin d'en faciliter l'accès et de rendre leur localisation plus rapide et plus efficace. Cela conduit aussi à l'exigence d'une gestion efficace des données vidéo qui ouvre la voie à de nouveaux domaines de recherche, tels que l'indexation et la recherche de vidéos qui respectent leur contenu spatio-temporel, visuels et sémantiques. Les systèmes d'indexation conçus ne sont pas encore assez performants car ils exploitent soit le texte ou le contenu. Cela donne la possibilité d'effectuer des recherches qui peuvent être basées sur le texte rattaché à la vidéo (contenu sémantique) ou basées sur le contenu visuel en utilisant ce qu'on appelle les caractéristiques de "bas niveau" (couleur, forme, texture). Par conséquent, ce processus d'indexation génère un problème majeur connu sous le nom de fossé sémantique. Ce dernier est défini comme une divergence entre la représentation de bas niveau d'une vidéo et son interprétation sémantique par les experts du domaine. Notre travail de thèse consiste en la conception et la réalisation d'un nouveau système d'indexation et de recherche des vidéos appelé OVIS (Ontology Video-Surveillance Indexing and retrieval System) et son expérimentation à l'aide de benchmarks universels de vidéos comme PETS 2012 et TRECVID 2016. La solution proposée est basée sur une ontologie du domaine de la vidéosurveillance qui respecte une convention de nommage syntaxique et utilise un ensemble de règles sémantiques SWRL (Semantic Web Rule Language).

Mots Clés :

Ontologie de vidéo surveillance; Indexation des vidéos; Recherche des vidéos; Evènements de foule; Fossé sémantique; Convention de nommage syntaxique; OVIS; SWRL; PETS 2012; TRECVID 2016.