Universitat
Autònoma
de Barcelona

Understanding Image Sequences: the Role of Ontologies in Cognitive Vision

A dissertation submitted by **Carles Fernández Tena** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor en Informàtica**.

Bellaterra, April 2010

Director: **Dr. Jordi Gonzàlez i Sabaté**
Centre de Visió per Computador
Dept. de Ciències de la Computació, Universitat Autònoma de Barcelona.
Co-director: **Dr. Xavier Roca i Marvà**
Dept. de Ciències de la Computació, Universitat Autònoma de Barcelona.
Centre de Visió per Computador

Centre de Visió
per Computador

This document was typeset by the author using $\LaTeX\,2_{\varepsilon}$.

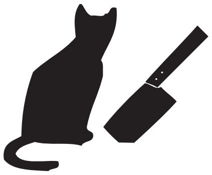
The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

ISBN 978-84-937261-2-6

Printed by Ediciones Gráficas Rey, S.L.

"Her taş, baş yarmaz."

"Bakmakla öğrenilseydi, kediler kasap olurdu. . . "



Acknowledgements

Aquesta tesi veu la llum fruit de molta, molta feina. Però especialment també és fruit de molts anys de col·laboració, treball en equip i gent fantàstica que m'ha fet fàcils els moments durs, i ha sabut alimentar la meva il·lusió dia rere dia. És el meu deure començar a agrair-vos tot això amb aquestes paraules.

Agraeixo l'oportunitat d'endinsar-me dins la recerca que fa uns quants anys em van oferir Jordi Gonzàlez, Juanjo Villanueva i Xavi Roca. Ells m'han donat carta blanca per a barallar-me amb dues de les meves curiositats més febrils, la visió per computador i la lingüística. No hauria cregut mai que m'hauria estat possible abraçar conjuntament ambdós camps, ha estat i encara és tota una aventura.

Durant aquests anys he vist crear-se i créixer a tot un model de grup. Danny, Ignasi, Mateu, ja migrats, han passat el relleu a un equip de gent genial que no tem donar passos de gegant. I would like to encourage those who are still hitting the road to persevere on our quest: Bhaskar, Marco, Murad, Ariel, Noha, Pep, Nataliya, Wenjuan, Miguel, Marc and Zhanwu. Andy, from now on I'll have to pay the beers, too…

Thanks to those who have put their words in my hands, clearing up their thick agendas just to help me conquer a new language: Ekain, Tuğçe, Noha, Marco, Jordi.

El CVC és i seguirà essent per mi una gran família, aquell lloc on fem i compartim, un engranatge de bon ambient i cordialitat, un entorn de feina sense igual. Vull agrair la confiança, gran disposició i companyerisme que he rebut de tots vosaltres: Helena, Josep, Raul, Jose Manuel, David, Enric…tants que no hi cabrien. Tampoc em puc oblidar de les grans persones del CVC que des de la seva feina diària han facilitat tantíssim la meva: Montse, Gigi, Anita, Raquels, Mireia, Mari, Joan, Pilar.

I will always treasure the stay that Pau and I had in Karlsruhe with Prof. H.–H. Nagel, Hanno, and Aleš. Vielen Dank für die weisen Ratschläge, interessante Gespräche und schöne Nachtessen alle zusammen. I also want to thank Dr. Reid, Eric, Nicola, and Ben for their kindness and attentions during my visits to Oxford. In this line, I keep great memories with the rest of members of the Hermes family: Dani, Preben, Thomas, and the others.

Vorrei anche ringraziare i miei amici di Firenze: Marco Bertini, Lorenzo, Gianluca, Beppone, Iacopo, Nicola, Andrea, Beppino, Marco Meoni, Fernando, Fede e tanti altri, che si sono presi cura di me durante il mio soggiorno. Grazie mille! Non potrò mai dimenticare la vostra ospitalità, i caffè macchiati, la finocchiona e quella bellissima bistecca!

A la Giraldilla, on cada divendres (o dijous, si festiu) ens reunim amb el Marçal, l'Alícia, l'Edu, el Mohammad i els altres, per endrapar entrepans amb pebrot, salsa del Betis, i bota de birra. Juntament amb en Ferran, en Joan Mas i el Javier, tots vosaltres, el meu despatx, sou la millor raó per anar a treballar cada dia amb un somriure.

Al Poal, el meu director de tesi i amic, que m'ha ensenyat tantes coses també fora del terreny acadèmic, i m'ha ofert motivació, paciència, confiança i suport incondicionals al llarg de tots aquests anys. Ell sempre sap mirar els conflictes des d'un punt de vista amable, i en tot moment ha vist la persona abans que el doctorand.

Als meus amics Pau i Ivan: el document que teniu a les mans no hauria estat possible sense els ànims i alegria constants que m'heu donat, al llarg de les pendents que hem anat passant tots junts. La *doctoramenta* del Pau m'ha empès molt fort per a aconseguir la meva, i espero que el meu intent tingui un efecte similar cap a l'Ivan. Ànims!!

Nando, Raúl, Pablo, Rau, vuestra amistad no tiene precio, con distancia o sin ella.

Ve burada Tuğçe'yi unutmmak istemiyorum, tabi ki! Senin herzaman desteğin ve bitmek bilmeyen aşkın, beni bu tezimi bitirmem de yardımcı oldu. Teşekkür ederim sevgilim, sensiz bir hayat düşünemiyorum.

Esta tesis está dedicada a mi familia, dado que ellos son la motivación última de mis esfuerzos. En especial, quiero dedicar el trabajo de todos estos años a la memoria de mi padre, cuyo ejemplo de vida marca mis pasos día tras día.

Abstract

The increasing ubiquitousness of digital information in our daily lives has positioned video as a favored information vehicle, and given rise to an astonishing generation of social media and surveillance footage. This raises a series of technological demands for automatic video understanding and management, which together with the compromising attentional limitations of human operators, have motivated the research community to guide its steps towards a better attainment of such capabilities. As a result, current trends on cognitive vision promise to recognize complex events and self-adapt to different environments, while managing and integrating several types of knowledge. Future directions suggest to reinforce the multi-modal fusion of information sources and the communication with end-users.

In this thesis we tackle the problem of recognizing and describing meaningful events in video sequences from different domains, and communicating the resulting knowledge to end-users by means of advanced interfaces for human–computer interaction. This problem is addressed by designing the high-level modules of a cognitive vision framework exploiting ontological knowledge. Ontologies allow us to define the relevant concepts in a domain and the relationships among them; we prove that the use of ontologies to organize, centralize, link, and reuse different types of knowledge is a key factor in the materialization of our objectives.

The proposed framework contributes to: (i) automatically learn the characteristics of different scenarios in a domain; (ii) reason about uncertain, incomplete, or vague information from visual –camera's– or linguistic –end-user's– inputs; (iii) derive plausible interpretations of complex events from basic spatiotemporal developments; (iv) facilitate natural interfaces that adapt to the needs of end-users, and allow them to communicate efficiently with the system at different levels of interaction; and finally, (v) find mechanisms to guide modeling processes, maintain and extend the resulting models, and to exploit multimodal resources synergically to enhance the former tasks.

We describe a holistic methodology to achieve these goals. First, the use of prior taxonomical knowledge is proved useful to guide MAP-MRF inference processes in the automatic identification of semantic regions, with independence of a particular scenario. Towards the recognition of complex video events, we combine fuzzy metric-temporal reasoning with SGTs, thus assessing high-level interpretations from spatiotemporal data. Here, ontological resources like T–Boxes, onomasticons, or factual databases become useful to derive video indexing and retrieval capabilities, and also to forward highlighted content to smart user interfaces. There, we explore the application of ontologies to discourse analysis and cognitive linguistic principles, or scene augmentation techniques towards advanced communication by means of natural language dialogs and synthetic visualizations. Ontologies become fundamental to coordinate, adapt, and reuse the different modules in the system.

The suitability of our ontological framework is demonstrated by a series of applications that especially benefit the field of smart video surveillance, viz. automatic generation of linguistic reports about the content of video sequences in multiple natural languages; content-based filtering and summarization of these reports; dialogue-based interfaces to query and browse video contents; automatic learning of semantic regions in a scenario; and tools to evaluate the performance of components and models in the system, via simulation and augmented reality.

Resum

La gran importància i omnipresència de la informació digital ha posicionat el vídeo com a vehicle preferent per a transmetre informació, i ha donat lloc a un espectacular creixement en la generació de multimèdia a les xarxes socials i de material de video vigilància. Aquesta situació exigeix tot un seguit de necessitats tecnològiques que han motivat moltes iniciatives de recerca per la millora en la comprensió automàtica del contingut en seqüències de vídeo. Com a resposta, la recerca en sistemes de visió cognitiva estudia sistemes capaços de reconèixer esdeveniments complexos i adaptar-se a diferents tipus d'entorn, tot i fent servir coneixement de diversa naturalesa.

En aquesta tesi ens proposem reconèixer i descriure el contingut de diferents situacions observades en seqüències de vídeo de diferents dominis, i comunicar la informació resultant a usuaris externs per mitjà d'interfícies d'interacció home–màquina avançades. Aquest problema s'aborda mitjançant el disseny dels mòduls d'alt nivell d'un sistema de visió cognitiva que empra models ontològics. Concretament, ens proposem: (i) fer que el sistema s'adapti a diferents escenaris d'un domini, i n'aprengui automàticament les característiques; (ii) que raoni sobre informació incerta, incompleta o imprecisa, tant de tipus visual (càmeres) com de tipus lingüístic (usuaris); (iii) que generi interpretacions sensates d'esdeveniments complexes a partir de l'anàlisi de dades espai-temps més bàsiques; (iv) que disposi d'interfícies de comunicació natural que puguin solventar les necessitats dels usuaris; i finalment, (v) trobar mecanismes que ens facilitin el disseny, manteniment i extensió dels models implicats, i formes de combinar sinèrgicament les tasques descrites.

Per tal d'avaluar de forma intel·ligent continguts de vídeo és necessari adoptar tècniques avançades de manipulació de la informació. La nostra aproximació opta per seguir els principis dels sistemes de visió cognitiva. Per a fer-ho, utilitzem processos d'aprenentatge basats en inferència MAP-MRF per a l'identificació de regions semàntiques en diferents escenaris; raonadors de lògica difusa i arbres de grafs de situació (SGTs) per a interpretar automàticament el contingut de vídeos; processos de pàrsing basats en representació del discurs i semàntica cognitiva per a implementar mòduls de comunicació lingüística; i tècniques de síntesi o augmentació d'escenes per a simulació i representació. Adicionalment, demostrem que l'ús d'ontologies per a organitzar, centralitzar, connectar i reutilitzar coneixement és un factor clau a l'hora de materialitzar els nostres objectius.

Els avantatges del sistema descrit es demostren amb un conjunt d'aplicacions que beneficien principalment el camp de la video vigilància, com ara: generació automàtica de descripcions en diverses llengües sobre el contingut de seqüències de vídeo; filtrat i resum d'aquests texts d'acord amb els seus continguts; interfícies de diàleg amb l'usuari que li permetin fer consultes i navegar pels continguts dels vídeos; aprenentatge automàtic de les regions semàntiques presents a un escenari; i eines per a avaluar el funcionament de diferents components i models del sistema, fent servir tècniques de simulació de comportaments i realitat augmentada.

Contents

List of Tables

List of Figures

Glossary and acronyms

cognitive vision system An artificial cognitive vision system evaluates image sequences from a recorded scene, and transforms the numerical results extracted from the evaluation into conceptual descriptions of temporal developments in the scene [94]. 12–14, 16, 17, 30, 34, 54, 59, 84, 89, 137, 140

DL *Description Logics.* A family of formal knowledge representation languages, which are more expressive than propositional logic but have more efficient decision problems than first-order predicate logic. They provide a logical formalism for Ontologies and the Semantic Web, and provide formal reasoning for terminological knowledge (T-Boxes) [9]. 34, 54, 81

DRS *Discourse Representation Structure.* Structure containing a semantic representation for cohesive linguistic units often larger than single sentences, e.g., multisentential passages, paragraphs, discourses, or texts [61]. 91–93, 95, 99

event modeling Process aiming to describe events of interest formally and to enable recognition of these events as they occur in video sequences [74]. 51, 52

FMTL *Fuzzy Metric Temporal horn Logic.* A form of logic in which conventional formalisms are extended by a temporal and a fuzzy component. The first one permits to represent, and reason about, propositions qualified in terms of time; the last one deals with uncertain or partial information, by allowing degrees of truth or falsehood [95, 112]. 28, 66, 68, 70, 71, 74, 126, 139

HSE *Human Sequence Evaluation.* Application domain of cognitive vision systems that specifically focus on the evaluation of image sequences involving human presence. Its modular scheme bases on the bidirectional communication between consecutive abstraction levels to transform image data into conceptual descriptions, and vice versa. 13, 14, 29, 140

lemma A lemma or citation form is the canonical form of a lexeme. Lemma refers to the particular form that is chosen by convention to represent the lexeme [89] (plural lemmata). 105

NL *Natural Language.* A language that is spoken, written, or signed by humans for general purpose communication, as distinguished from computer programming languages or those used in the study of formal logic. Linguistically, NL only applies to a language that has evolved naturally, and its study primarily involves native, first language speakers [127]. 17, 31, 33, 52, 65, 74, 78, 84, 88, 93, 102, 103, 108, 110, 116–121, 124, 127, 134, 139, 147, 149, 153, 155–157

NLG *Natural Language text Generation.* The process of generating natural language text from a machine representation system such as a knowledge base or a logical formula. 87–90, 92, 93, 96, 98, 116, 121, 147

NLU *Natural Language text Understanding.* The process of deciding for the most appropriate interpretation of a natural language textual input, out from a set of interpretations in form of logical formulae. 87–90, 98, 103–107, 110, 111, 121, 126, 137, 138, 147

ontology Formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to describe the domain [47]. 9, 16–18, 21, 29, 31, 34, 51, 52, 54–56, 60, 61, 65, 72, 74, 77, 78, 84, 89, 103–105, 121, 133, 134, 138, 140

REG *Referring Expression Generation.* Task of deciding which expressions should be used to refer to entities. 93, 95, 98, 100–103, 114, 147, 148, 150–152, 154–156

SGT *Situation Graph Tree.* Deterministic model that explicitly represents and combines the specialization, temporal, and semantic relationships of its constituent conceptual predicates. SGTs are used to describe the behavior of an agent in terms of situations he can be in [5, 7]. 28–30, 32, 60–66, 70, 74, 77, 84, 89, 110, 112–114, 127, 129–131, 137, 139

trajectory In the field of behavior understanding, series of positions of a moving object over time, from the moment it enters to the moment it exits a scene. 25–27, 38, 41, 44–47, 56, 67, 79, 81, 82, 112, 128, 137

video understanding Field devoted to translate low-level content in video sequences into high-level semantic concepts [74]. 8, 21, 31, 51, 52, 140

Chapter 1

Introduction

"In 2007, for the first time ever, more information was generated in one year than had been produced in the entire previous five thousand years —the period since the invention of writing."

Me the media: rise of the conversation society (2009),
by J. Bloem, M. van Doorn, and S. Duivestein

The revolution of information experienced by the world in the last century, especially emphasized by the household use of computers after the 1970s, has led to what is known today as the society of knowledge. Digital technologies have converted postmodern society into an entity in which networked communication and information management have become crucial for social, political, and economic practices. The major expansion in this sense has been rendered by the global effect of the Internet: since its birth, it has grown into a medium that is uniquely capable of integrating modes of communication and forms of content.

In this context, the assessment of interactive and broadcasting services has spread and generalized in the last decade –e.g., residential access to the Internet, video-on-demand technologies–, posing *video* as the privileged information vehicle of our time, and promising a wide variety of applications that aim at its efficient exploitation. Today, the automated analysis of video resources is not tomorrow's duty anymore. The world produces a massive amount of digital video files every passing minute, particularly in the fields of multimedia and surveillance, which open windows of opportunity for smart systems as vast archives of recordings constantly grow.

Automatic content-based video indexing has been requested for digital multimedia databases for the last two decades [39]. This task consists of extracting high-level descriptors that help us to automatically annotate the semantic content in video sequences; the generation of reasonable semantic indexes makes it possible to create powerful engines to search and retrieve video content, which finds immediate applications in many areas: from the efficient access to digital libraries to the preservation and maintenance of digital heritage. Other usages in the multimedia domain would also include virtual commentators, which could describe, analyze, and summarize the development of sport events, for instance.

More recently, the same requirements have applied also to the field of video surveillance. Human operators have attentional limitations that discourage their involvement in a series of tasks that could compromise security or safety. In addition, surveillance systems have strong storage and computer power requirements, deal with continuous 24/7 monitoring, and manage a type of content that is susceptible to be highly compressed. Furthermore, the number of security cameras increases exponentially worldwide on a daily basis, producing huge amounts of video recordings that may require further supervision. The conjunction of these facts establishes a need to automatize the visual recognition of events and content-based forensic analysis on video footage.

We find a wide range of applications coming from the surveillance domain that point to real-life, daily problems: for example, a smart monitoring of elder or disabled people makes it possible to recognize alarming situations, and speed up reactions towards early assistance; road traffic surveillance can be useful to send alerts of congestion or automatically detect accidents or abnormal occurrences; similar usage can be directed to urban planning, optimization of resources for transportation allocations, or detection of abnormality in crowded locations –airports, lobbies, etc.–.

Such a vast spectrum of social, cultural, commercial, and technological demands have repeatedly motivated the research community to direct their steps towards a better attainment of video understanding capabilities.

1.1 Collaborative efforts on video event understanding

A notable amount of EU research projects have been recently devoted to the unsupervised analysis of video contents, in order to automatically extract events and behaviors of interest, and interpret them in selected contexts. These projects measure the pulse of the research in this field, demonstrate previous success on particular initiatives, and propose a series of interesting applications to such techniques. And, last but not least, they motivate the continuation of this line of work. Some of them are briefly described next, and shown in Figs 1.1 and 1.2.

- *ADVISOR* [1] (IST-11287, 2000–2002). It addresses the development of management systems for networks of metro operators. It uses CCTV for computer-assisted automatic incident detection, content based annotation of video recordings, behavior pattern analysis of crowds and individuals, and ergonomic human computer interfaces.

- *ICONS* [2] (DTI/EPSRC LINK, 2001–2003). Its aim is to advance towards (i) zero motion detection, detection of medium- to long-term visual changes in a scene –e.g., deployment of a parcel bomb, theft of a precious item–, and (ii) behavior recognition –characterize and detect undesirable behavior in video data, such as thefts or violence– only from the appearance of pixels.

- *AVITRACK* [3] (AST-CT-502818, 2004–2006). It develops a framework for auto-

[1] http://www-sop.inria.fr/orion/advisor/

[2] http://www.dcs.qmul.ac.uk/research/vision/projects/icons/

[3] http://www.avitrack.net/

Figure 1.1: Snapshots of the referred projects. (a) *AVITRACK*, (b) *ADVISOR*, (c) *BEWARE*, (d) *VIDI-Video*, (e) *CARETAKER*, (f) *ICONS*, (g) *ETISEO*, (h) *HERMES*.

matically supervision of commercial aircraft servicing operations from the arrival to the departure on an airport's apron. A prototype for scene understanding and simulation of the apron's activity was going to be implemented during the project on Toulouse airport.

- *ETISEO* [4] (Techno-Vision, 2005–2007). It seeks to work out a new structure contributing to an increase in the evaluation of video scene understanding. ETISEO focuses on the treatment and interpretation of videos involving pedestrians and (or) vehicles, indoors or outdoors, obtained from fixed cameras.

- *CARETAKER* [5] (IST-027231, 2006–2008). This project aims at studying, developing and assessing multimedia knowledge-based content analysis, knowledge extraction components, and metadata management sub-systems in the context of automated situation awareness, diagnosis and decision support.

- *SHARE* [6] (IST-027694, 2006–2008). It offers an information and communication system to support emergency teams during large-scale rescue operations and disaster management, by exploiting multimodal data –audio, video, texts, graphics, location–. It incorporates domain dependent ontology modules, and allows for video/voice analysis, indexing/retrieval, and multimodal dialogues.

- *HERMES* [7] (IST-027110, 2006–2009). Extraction of descriptions of people's behavior from videos in restricted discourse domains, such as inter-city roads, train stations, or lobbies. The project studies human movements and behaviors at several scales –agent, body, face–, and the final communication of meaningful contents to end-users.

- *BEWARE* [8] (EP/E028594/1, 2007–2010). The project aims to analyze and combine data from alarm panels and systems, fence detectors, security cameras, public sources and even police files, to unravel patterns and signal anomalies, e.g., by making comparisons with historical data. BEWARE is self-learning and suggests improvements to optimize security.

- *VIDI-Video* [9] (IST-045547, 2007–2010). Implementation of an audio-visual semantic search engine to enhance access to video, by developing a 1000 element thesaurus to index video content. Several applications have been suggested in surveillance, conferencing, event reconstruction, diaries, and cultural heritage documentaries.

- *SAMURAI* [10] (IST-217899, 2008–2011). It develops an intelligent surveillance system for monitoring of critical public infrastructure sites. It is to fuse data from networked heterogeneous sensors rather than using CCTV alone; to develop real-adaptative behavior profiling and abnormality detection, instead of

[4] http://www-sop.inria.fr/orion/etiseo/
[5] http://www.ist-caretaker.org/
[6] http://www.ist-share.org/
[7] http://www.hermes-project.eu/
[8] http://www.dcs.qmul.ac.uk/sgg/beware/
[9] http://www.vidi-video.it/
[10] http://www.samurai-eu.org/

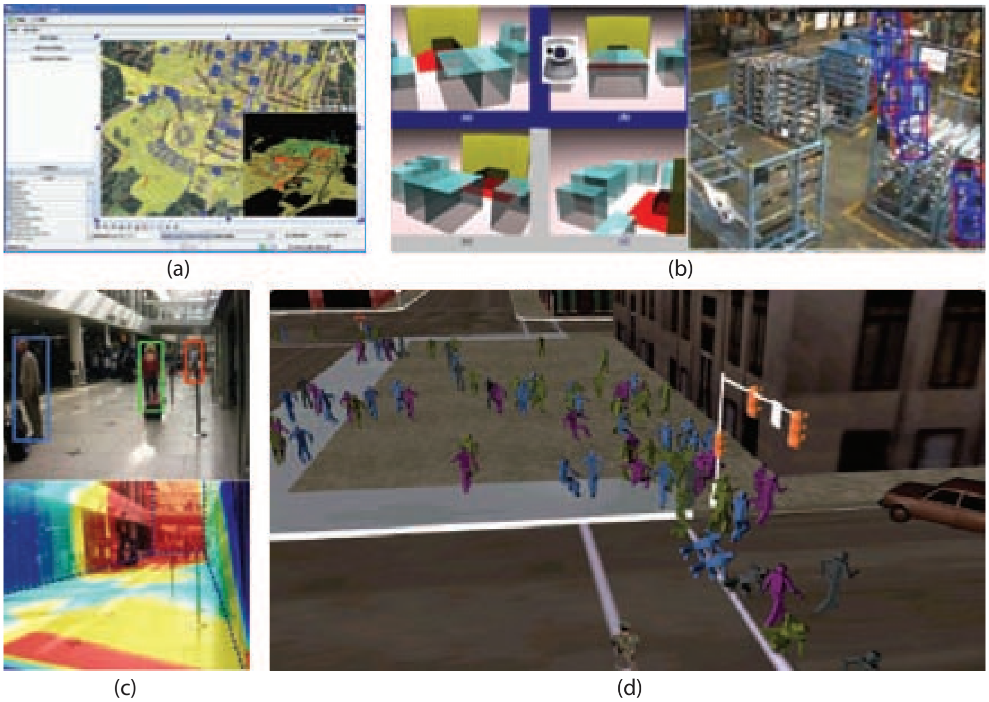


Figure 1.2: Some of the most recent projects in the field. (a) *SHARE*, (b) *SCOVIS*, (c) *SAMURAI*, (d) *ViCoMo*.

using predefined hard rules; and to take command input from human operators and mobile sensory input from patrols, for hybrid context-aware behavior recognition.

- *SCOVIS* [11] (IST-216465, 2007–2013). It aims at automatic behavior detection and visual learning of procedures, in manufacturing and public infrastructures. Its synergistic approach based on complex camera networks also achieves model adaptation and camera network coordination. User's interaction improves behavior detection and guides the modeling process, through high-level feedback mechanisms.

- *ViCoMo* [12] (ITEA2-08009, 2009–2012). This project concerns advanced video-interpretation algorithms on video data that are typically acquired with multiple cameras. It is focusing on the construction of realistic context models to improve the decision making of complex vision systems and to produce a faithful and meaningful behavior.

As it can be seen, many efforts have been taken in the last decade, and are still increasing nowadays, in order to tackle the problem of video interpretation and intel-

[11] http://www.scovis.eu/

[12] http://www.vicomo.org/

ligent video content management. It is clear from this selection that current trends on the field suggest a tendency to focus on the multi-modal fusion of different sources of information, and on more powerful communication with end-users. From the large amount of projects existing in the field we derive another conclusion: such a task is not trivial at all, and requires research efforts from many different areas to be joint into *collaborative approaches*, which success where individual efforts fail.

In this thesis we tackle the problem of recognizing and describing meaningful events in video sequences, and communicating the resulting knowledge to end-users by means of advanced user interfaces. This will be done particularly for the field of video surveillance, although many of the results that will be presented –e.g., query understanding based on natural language, automatic indexing of video events– can be also applied to multimedia applications. The series of challenges coming from these applications will be addressed by designing the high-level modules of a cognitive vision framework exploiting ontological knowledge.

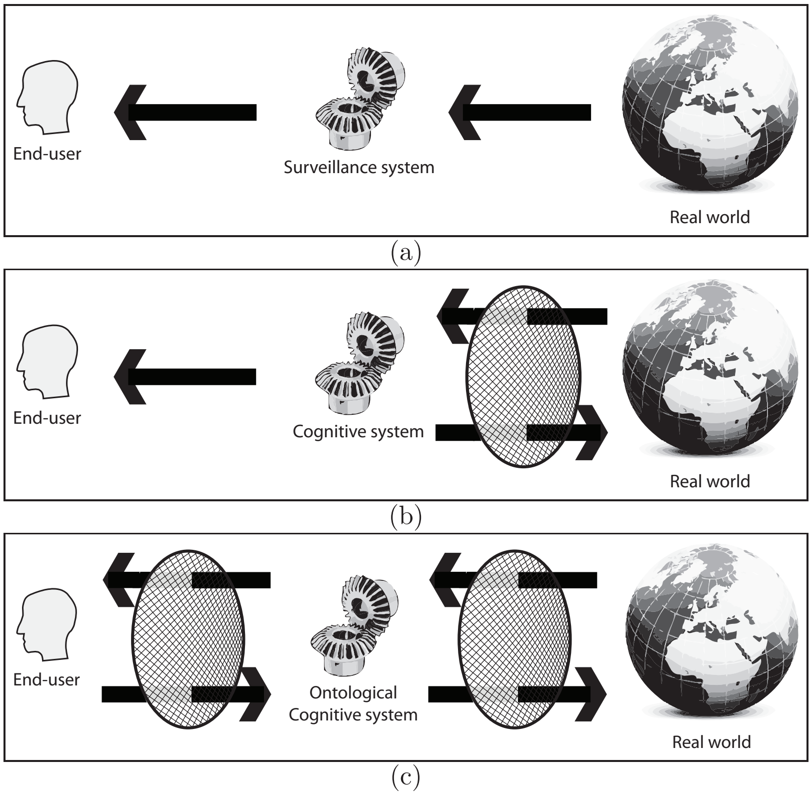
1.2 Past, present, and future of video surveillance

The field of video surveillance has experienced a remarkable evolution in the last decades, which can help us think of the future characteristics that would be desirable for it. In the traditional video surveillance scheme, the primary goal of the camera system was to present to human operators more and more visual information about monitored environments, see Fig. 1.3(a). First-generation systems were completely passive, thus having this information entirely processed by human operators. Nevertheless, a saturation effect appears as the information availability increases, causing a decrease in the level of attention of the operator, who is ultimately in charge of deciding about the surveilled situations.

The following generation of video surveillance systems used digital computing and communications technologies to change the design of the original architecture, customizing it according to the requirements of the end-users. A series of technical advantages allowed them to better satisfy the demands from industry, i.e., higher-resolution cameras, longer retention of recorded video –DVRs replaced VCRs, video encoding standards appeared–, reduction of costs and size, remote monitoring capabilities provided by network cameras, or more built-in intelligence, among others [98].

The continued increase of machine intelligence has derived into a new generation of smart surveillance systems lately. Recent trends on computer vision and artificial intelligence have deepened into the study of cognitive vision systems, which use visual information to facilitate a series of tasks on sensing, understanding, reaction, and communication, see Fig. 1.3(b). Such systems enable traditional surveillance applications to greatly enhance their functionalities by incorporating methods for:

- Recognition and categorization of objects, structures, and events;
- Learning and adaptation to different environments;
- Representation, memorization, and fusion of various types of knowledge;
- Automatic control and attention.



(a)

(b)

(c)

Figure 1.3: Evolution of video surveillance systems, since its initial passive architecture (a) to the reactive, bidirectional communication scheme offered by cognitive vision systems (b), which highlight relevant footage contents. By incorporating ontological and interactive capabilities to this framework (c), the system performs like a *semantic filter* also to the end-users, governing the interactions with them in order to adapt to their interests and maximize the efficiency of the communication.

As a consequence, the relation of the system with the world and the end-users is enriched by a series of sensors and actuators –e.g., distributions of static and active cameras, enhanced user interfaces–, thus establishing a bidirectional communication flow, and closing loops at a sensing and semantic level. The resulting systems provide a series of novel applications with respect to traditional systems, like automated video commentary and annotation, or image-based search engines. In the last years, European projects like `CogVis` [13] or `CogViSys` [14] have investigated these and other potential applications of cognitive vision systems, especially concerning video surveillance.

Recently, a paradigm has been specifically proposed for the design of cognitive vision systems aiming to analyze human developments recorded in image sequences.

[13] http://www.comp.leeds.ac.uk/vision/cogvis/

[14] http://cogvisys.iaks.uni-karlsruhe.de/

This is known as Human Sequence Evaluation (HSE) [43]. An HSE system is built upon a linear multilevel architecture, in which each module tackles a specific abstraction level. Two consecutive modules hold a bidirectional communication scheme, in order to (i) generate higher-level descriptions based on lower-level analysis –bottom-up inference–, and (ii) support low-level processing with high-level guidance –top-down reactions–. HSE follows as well the aforementioned characteristics of cognitive vision systems.

Nonetheless, although cognitive vision systems conduct a large number of tasks and success in a wide range of applications, in most cases the resulting prototypes are tailored to specific needs or restricted to definite domains. Hence, current research aims to increase aspects like *extensibility*, *personalization*, *adaptability*, *interactivity*, and *multi-purpose* of these systems. In particular, it is becoming of especial importance to stress the paper of communication with end-users in the global picture, both for the fields of surveillance and multimedia: end-users should be allowed to automatize a series of tasks requiring content-mining, and should be presented the analyzed information in a suitable and efficient manner, see Fig. 1.3(c).

As a result of these considerations, the list of objectives to be tackled and solved by a cognitive vision system has elaborated on the original approach, which aimed at the single –although still ambitious today– task of transducing images to semantics. Nowadays, the user itself has become a piece of the puzzle, and therefore has to be considered a part of the problem.

1.3 *Mind the gaps*

The search and extraction of meaningful information from video sequences is dominated by 5 major challenges, all of them defined by *gaps* [116]. These gaps are disagreements between the real data and that one expected, intended, or retrieved by any computer-based process involved in the information flow conducted between the acquisition of data from the real world, and until its final presentation to the end-users. The 5 gaps are presented next, see Fig 1.4(a).

Sensory gap The gap between an object in the world and the information in an image recording of that scene. All these recordings will be different due to variations in viewpoint, lighting, and other circumstantial conditions.

Semantic gap The lack of coincidence between the information that one can extract from the sensory data and the interpretation that same data has for a user in a given situation. It can be understood as the difference between a visual concept and its linguistic representation.

Model gap The impossibility to theoretically account the amount of notions in the world, due to the limited capacity to learn them.

Query/context gap The gap between the specific need for information of an end-user and the possible retrieval solutions manageable by the system.

Interface gap The limited scope of information that a system interface offers, compared to the amount of data actually intended to transmit.

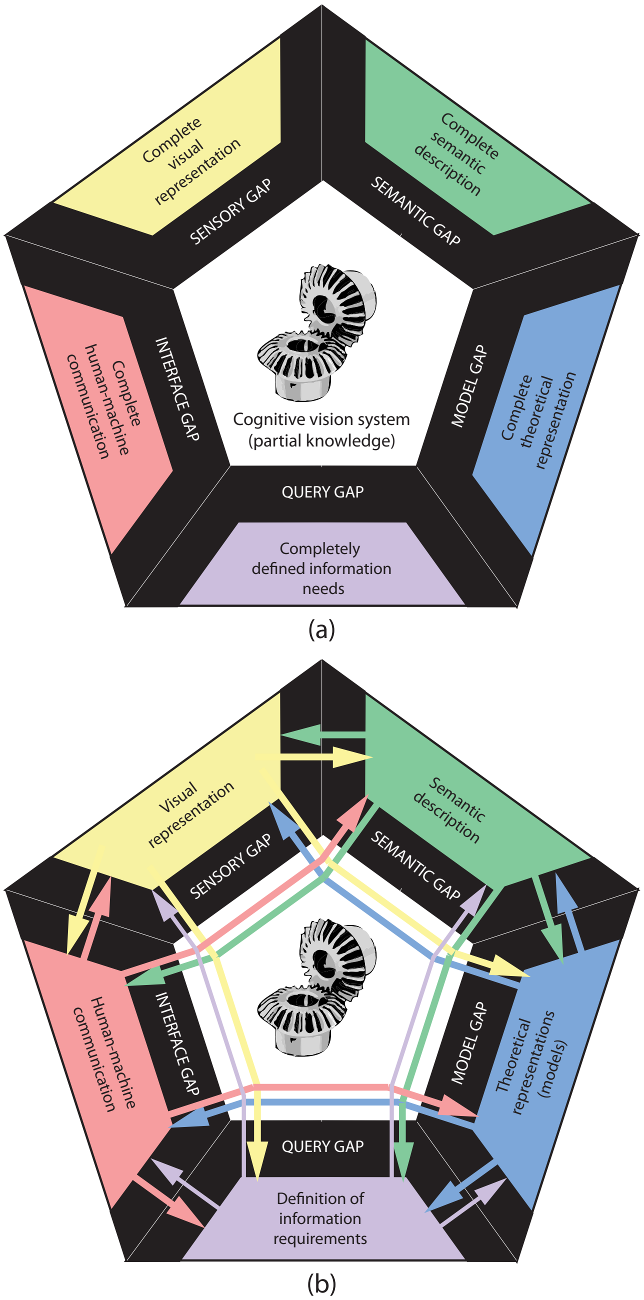

Complete visual representation
Complete semantic description
SENSORY GAP
SEMANTIC GAP
INTERFACE GAP
MODEL GAP
Complete human-machine communication
Cognitive vision system (partial knowledge)
Complete theoretical representation
QUERY GAP
Completely defined information needs
(a)
Visual representation
Semantic description
SENSORY GAP
SEMANTIC GAP
Human-machine communication
INTERFACE GAP
MODEL GAP
Theoretical representations (models)
QUERY GAP
Definition of information requirements
(b)

Figure 1.4: (a) The five gaps that need to be bridged for the successful analysis, extraction, search, retrieval, and presentation of video content. (b) In some cases, a collaborative and integrative use of different knowledge sources allows us to achieve or enrich the accomplishment of these tasks. Arrows stand for reusing ontological knowledge to enhance analyses in other areas.

Although each of these challenges becomes certainly difficult to overcome by its own, a proper centralization of information sources and the wise reutilization of knowledge derived from them facilitates the overwhelming task of bridging each of these gaps. There exist multiple examples of how the multiple resources of the system can be redirected to solve problems in a different domain, let us consider three of them:

- From *semantic* to *sensory* gap: tracking errors or occlusions at a visual level can be identified by high-level modules that imply semantics oriented to that end. This way, the system can be aware of where and when a target is occluded, and predict its reapparition.
- From *sensory* to *interface* gap: the reports or responses in user interfaces can become more expressive by adding selected, semantically relevant key-frames from the sensed data.
- From *interface* to *query* gap: in case of syntactic ambiguities in a query –e.g., "*zoom in on any person in the group that is running*"–, end-users can be asked about their real interests via a dialogue interface: "*Did you mean 'the group that is running', or 'the person that is running'?*".

Given the varied nature of types of knowledge involved in our intended system, an ontological framework becomes a sensible choice of design: such a framework integrates different sources of information by means of temporal and multi-modal fusion –horizontal integration–, using bottom-up or top-down approaches –vertical integration–, and incorporating prior hierarchical knowledge by means of an extensible ontology.

We propose the use of ontologies to help us integrate, centralize, and relate the different knowledge representations –visual, semantic, linguistic, etc.– implied by the different modules of the cognitive system. By doing so, the relevant knowledge or capabilities in a specific area can be used to enhance the performance of the system in other distinct areas, as represented in Fig. 1.4(b). Ontologies will enable us to formalize, account, and redirect the semantic assets of the system in a given situation, and exploit them to empower the aforementioned capabilities, especially targeting the possibilities of interaction with end-users.

1.4 Thesis scope and contributions

This thesis describes a complete framework for the high-level modules of an artificial cognitive vision system; particularly, this framework is devoted to ontologically-based cognitive video surveillance. The work done through the different chapters pursues three major lines of contribution:

- **High-level interpretation** of complex human and vehicle behaviors, in real scenes of different domains.
- Establishment of natural and effective channels of **advanced interaction** with end-users, regarding the communication of video contents.

- Development of an **ontological framework** to guide the top-down modeling of the expert databases. This framework centralizes the multiple types of knowledge involved by the system –*visual*, *conceptual*, *linguistic*–, and facilitates their integration into a large number of applications.

In addition, next table presents a summarized account of the specific tasks achieved by the ontological cognitive surveillance framework presented in this thesis, and the chapters in which these contributions appear:

Location	Contributions
Chapter 3	– Semantic region learning
Chapter 4	– Ontology-based top-down modeling for video understanding, – Interpretation/indexing of video events and behaviors, – Content filtering and episodical segmentation of videos
Chapter 5	– Generation of multilingual NL descriptions of videos, – Summarization / selection of contents, – Recognition of NL linguistic input / query retrieval, – Supervised linguistic rule learning, – Authoring tools, – Model-based simulation, – Component performance evaluation

This thesis is organized following the distribution of modules shown in Fig 1.5. Next chapter reviews recent literature on the recognition of semantic events and behaviors in video sequences, especially considering work related to cognitive vision systems, ontologies, and advanced user interfacing for surveillance applications. The accounting of semantic properties for the locations where the events occur varies for every new scene; for this reason, **Chapter 3** proposes a methodology to automatically learn meaningful semantic regions in scenarios of a given domain. **Chapter 4** suggests the reader a methodology to build the different semantic models described, including the ontology, and explains how to apply them to achieve efficient reasoning and understanding of tracked visual information. **Chapter 5** describes the three modules used by the system to provide capabilities of advanced interaction and communication with end-users: generation of textual descriptions, understanding of user queries, and generation or augmentation of synthetically animated scenes. Finally, **Chapter 6** briefly reviews the topics discussed in the different sections of this thesis, and, as a conclusion, establishes future lines of work that could eventually fix the current issues of the presented framework.

Resum

La societat actual s'ha vist enormement influenciada per les tecnologies digitals en els darrers anys. Avui en dia, l'ingent producció de seqüències de vídeo de tipus molt divers (grabacions de vigilància, producció audiovisual, multimèdia de caire social)

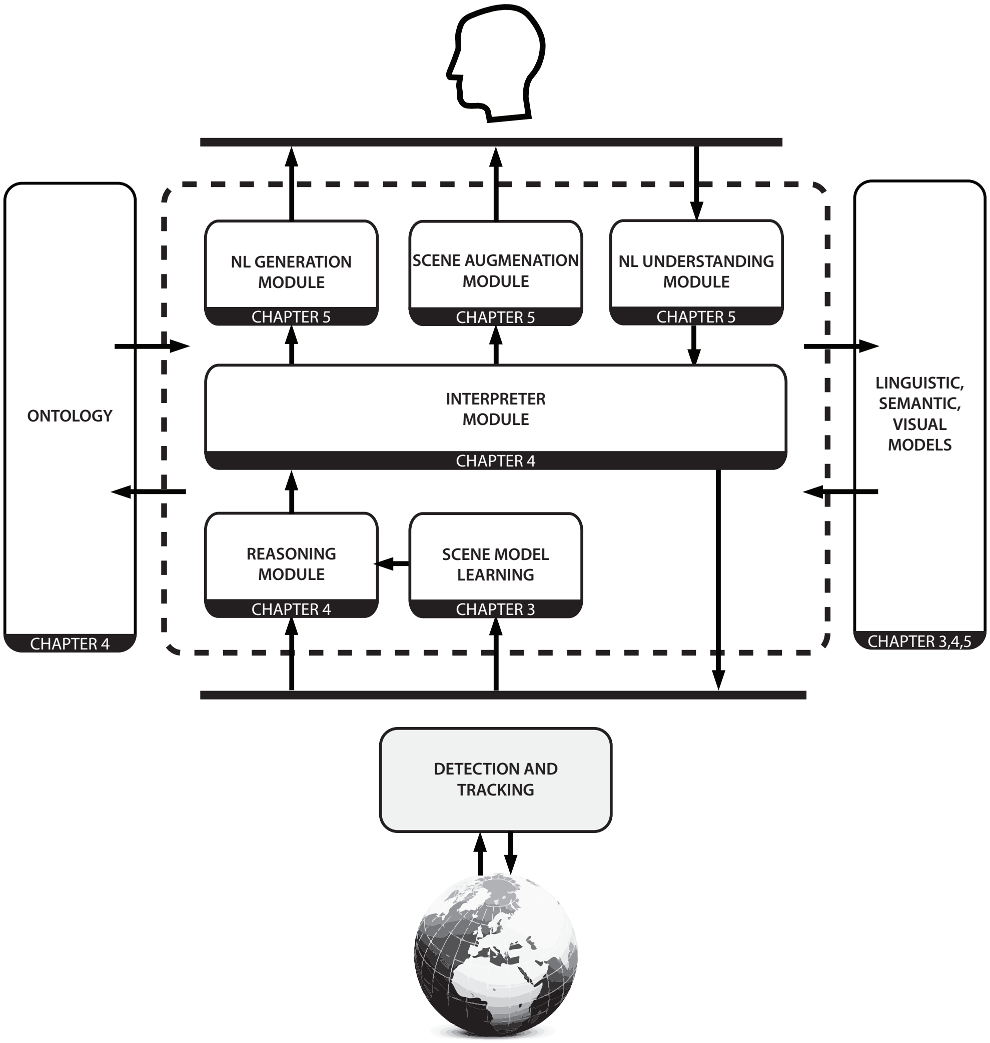


Figure 1.5: General architecture of the ontology-based cognitive vision framework proposed in this thesis. The implied modules are distributed along the different chapters of the book.

exigeix millores tecnològiques per a l'exploració automàtica, categorització, indexació i cerca de vídeos en quant al seu contingut semàntic. La gran quantitat de projectes europeus dedicats a perseverar en aquest objectiu esdevé una clara senyal de la posició d'importància que aquest camp ocupa dins les noves tecnologies de la informació.

Els sistemes de vídeo vigilància són un exemple molt clar de com s'ha anat produint una clara evolució en relació a l'anàlisi de continguts de vídeo. Als primers sistemes, tota la feina recau sobre operaris humans, que inspeccionen visualment la totalitat del metratge. Els nous avenços en detecció i seguiment visual permeten la incorporació de tècniques més sofisticades que guien l'atenció de l'usuari final, facilitant-li la identificació d'activitats concretes. Avui en dia, la implantació de sistemes visuals cognitius possibilita tasques de reconeixement d'objectes i situacions, control automàtic i aprenentatge continu. Quan es parla del següent pas de l'evolució, sembla que s'està d'acord en enfortir la relació amb els usuaris, per mitjà d'interfícies intel·ligents que filtrin les necessitats específiques dels usuaris de forma eficient i natural.

Per tal d'aconseguir aquesta fita, s'ha de superar tot un seguit d'incompatibilitats tradicionalment presents en sistemes d'aquestes característiques: les representacions visuals obtingudes contenen informació poc precisa de la realitat; les interpretacions que fa un sistema sobre una situació s'allunyen de les què faria una persona; no ens és possible modelar tot allò que ens pot interessar reconèixer; el sistema pot no entendre correctament les necessitats específiques d'un usuari; i les dades proporcionades pel sistema sempre es veuran limitades per la interfície, essent una reducció dràstica de tot el coneixement involucrat en la resolució de la tasca.

Per tal de reduir progressivament totes aquestes dificultats, aquesta tesi proposa l'arquitectura d'alt nivell d'un sistema de visió cognitiva artificial. Es para especial atenció en el disseny de recursos ontològics, que permeten una millor organització i centralització d'informació de diferent naturalesa. D'altra banda, un altre aspecte clau és el disseny de mòduls per a establir comunicació d'alt nivell amb l'usuari, permetent així aplicacions tals com interfícies de diàleg i consulta en múltiples llengües, simulació i avaluació de components, o selecció i resum automàtic de continguts, entre altres.

Chapter 2

Related work

> *"No reference is truly direct – every reference depends on some kind of coding scheme. It's just a question of how implicit it is."*
>
> *Gödel, Escher, Bach: An eternal golden braid* (1979), by D.R. Hofstadter

The field of video understanding has been tackled by the research community for many years now, deriving a large amount of techniques for event recognition based on both statistical and model-based approaches. This chapter reviews part of this work: we examine the most common terminologies for the organization of events, trajectory-based methods for activity recognition and semantic region learning, frequent probabilistic and symbolic tools used for video event recognition, and extensions for user interaction dealing with natural language and virtual environments.

There is an impressive collection of literary works dedicated to the understanding of content in video sequences, and its further communication to end-users. Efforts to survey this area of research have been unceasing for the last fifteen years [2, 26, 55, 101, 125, 74]. In order to introduce the field, in Section 2.1 we initially consider some basic ideas regarding the organization and representation of knowledge, and especially, the semantic classification of video events traditionally used for video understanding.

From there, the selection of references compiled in this chapter follows the distribution of chapters of the thesis: Section 2.2 reviews research on automatic *learning of semantic regions* from video surveillance sequences. After that, Section 2.3 presents a representative selection of works on the prolific field of *event/activity/behavior recognition* in video sequences, summarizing the many approaches and techniques that have been used in this field for the last decade, and justifying our decision to use symbolic models. Section 2.4 considers the work done on advanced user interaction, especially focusing on Natural Language (NL) interfaces and virtual or augmented reality. Finally, Section 2.5 reviews the general use of ontologies to interrelate visual, semantic, and linguistic knowledge, and how ontological knowledge can benefit applications aiming for advanced means of user interaction.

2.1 Knowledge representation: statistical vs model-based

As described in [68], two main types of approaches for knowledge acquisition and representation have been identified in the literature: implicit, accomplished by machine learning methods, and explicit, using model-based approaches. The former have proved to be robust for discovering complex dependencies between low-level image data and perceptually higher level concepts, and can also handle high dimensionality issues. On the other hand, model-based reasoning approaches use prior knowledge in form of explicitly defined facts, models, and rules, thus providing coherent semantics for a specific context.

The model-based reasoning approach uses predefined semantic models to anticipate events or behaviors that come associated to certain locations in the scenario, e.g. waiting in a bus stop or sitting on a table. For this reason, this approach is found especially useful for (i) applications aiming to very specific or regulated contexts, and (ii) those requiring to deal with a precise set of unlikely but possible situations. This is the case for the fields of surveillance and sports, for example.

Table 2.1 summarizes the main characteristics of these two approaches, statistical learning and model-based reasoning. The main features listed for each approach have been classified as advantages (✓) or disadvantages (✕) for a rapid exploration.

Semantic organization and terminologies

Words like *behavior*, *activity*, *action*, *scenario*, *gesture*, *primitive* or *event* are often used in the literature to designate the same idea, although with slight variations. Occurrences in a video sequence are categorized by each author according to their complexity, but usually from different perspectives, which leads to controversy and ambiguity in their use. Next, we compile a small list of semantic hierarchies that are often cited in literature, and establish the meaning of the terms that we will be using during the following chapters.

Table 2.2 gives examples to the terminologies discussed next. In the first place, Nagel [93] organizes occurrences semantically into *change*, *event*, *verb*, *episode* and *history*, sorted increasingly by semantic complexity and temporal scope. An episode is a collection of sequential verbs, and a history often involves goals. Bobick [20] differentiates among *movement*, *activity* –when the movement follows a cyclic pattern, such as walking–, and *action* –for more complex events like entering a location–. Hongeng and Nevatia's terminology [52] builds upon *simple/complex events*, which are additionally classified as happening one at a time (*single thread*) or many at the same time (*multiple thread*). In the case of Park and Aggarwal [103], the complexity is given by the scope of membership, dividing events into *pose*, *gesture*, *action*, and *interaction*. Gonzàlez [44] considers sequences of *actions* as *activities*, and proposes *behaviors* as the interpretation of activities in given contexts. Most researchers use minimal variations over any of the previous organizations, e.g., Brémond [24] adapts the hierarchical event categories in [52].

The terminology used in this thesis is based on that proposed by Gonzàlez [44] and rethought in terms of ontological organization. As it will be explored in Chapter 4, the first conceptual level represents basic spatiotemporal information –walk, run, bend–

Statistical learning

- ✓ Models are learned algorithmically in an automatic fashion. The role of the experts is reduced to provide consisting samples for training, and in some cases supervise the process.
- ✓ Models are also easily updated, by just providing additional or new training samples.
- ✗ The correctness of the models relies on how representative are the training samples of the targeted domain. An accurate selection of training material may be necessary.
- ✗ Rare or uncommon events are hardly learned by observation, given the huge casuistry of possible developments (e.g., identifying *violence*, *thefts*). A common limitation in these cases is to detect only abnormal occurrences, i.e. those that fall out of the learned statistics.
- ✗ Aiming the models towards specific applications may require very precise training, which in some cases is more expensive and less robust than modeling by hand.

Model-based reasoning

- ✓ The domain of interest is precisely defined, and becomes fully controllable.
- ✓ As a consequence, those rules and configurations found relevant by experts are usually less susceptible to fail.
- ✓ Certain complex semantics that are difficult to learn may be easy to model. For instance, it is not straightforward to learn that a stranger accessing a computer represents a security risk, but it is easy to model such improbable –but assumed– occurrence.
- ✗ Models have to be manually defined by experts, in contrast to those automatically learned by statistical approaches.
- ✗ Content to include has to be carefully evaluated and formally described.
- ✗ It is desirable for knowledge bases to be still suitable for future applications, but this may not be the case for model-based reasoning. Relevant data may lack persistence, e.g., typewriters can be manually modeled and progressively fall into disuse in regular contexts.

Table 2.1

A comparison between characteristics of the two main approaches for knowledge acquisition and representation: statistical learning (implicit) and model-based reasoning (explicit).

Nagel (1988)

Change	Event	Verb	Episode	History
(Motion)	*Moving slowly*	*Driving by a road*	*Overtaking another car*	*Exiting a gas station*

Bobick (1997)

Movement	Activity	Action
(Motion)	*Walking*	*Entering a location*

Hongeng and Nevatia (2001)

Simple event	Complex event
approach a person	*converse (approach a person + stay around)*

Park and Aggarwal (2003)

Pose	Gesture	Action	Interaction
(Motion)	*Moving arm*	*Shaking hands*	*Greeting someone*

Gonzàlez (2004)

Movement	Action	Activity	Behavior
(Motion)	*Walking, bending*	*Approaching, chasing*	*Stealing an object*

Our proposal

Pose	Status	Contextualization	Interpretation
(Motion)	*Bending*	*Picking up something, somewhere*	*Stealing an object*

Table 2.2
Most common event terminologies, and our proposal.

. A second level contextualizes different observed entities, establishing links among them in order to validate schemes of multi-part events –leave something somewhere, enter a location, meet someone somewhere–. Finally, the situation of these events in specific temporal/semantic contexts will permit us to suggest high-level interpretations of behaviors.

2.2 Learning semantic regions

Detecting high-level events and behaviors in dynamic scenes requires to interpret "semantically meaningful object actions" [33], which in the concrete case of urban video surveillance restricts to monitoring and evaluating human and traffic activities in wide or far-field outdoor scenarios. Under such conditions, current state-of-the-art approaches infer activity descriptions mainly based on observed or expected motion within *regions of semantic relevance*. Therefore, our task demands an explicit description of locations in the scenario of interest, in terms of a series of semantic characteristics that can be found or anticipated in these zones.

In the literature, semantic regions can be either provided beforehand [95, 43, 40, 84] or automatically computed from static or dynamic features of the scene [92, 132, 140]. In the latter case, techniques for the automatic learning of semantic regions are commonly based on observed trajectories [131, 85] rather than on the appearance of pixels, given that appearances are usually view-variant, scene-dependant, and require considerable computational effort, thus being inconvenient for surveillance.

On the other hand, trajectories, understood as the series of positions of an object over time, from entering to exiting a scene, are considered by most authors as the most useful information to embed the behavior of moving objects [141]. Extensive work has been done on behavior understanding based on trajectory analysis: early results on motion-based behavior analysis were reported in [60], where spatial distributions of trajectories were modeled by training two competitive neural networks using vector quantization. Most often, trajectory clustering has been accomplished via mixtures of Gaussians: in [120], Stauffer and Grimson develop a tracking system that detects unusual events regarding their size or direction. The system was tested on roads and pedestrian paths. Hu *et al.* [56] described a generic framework to automatically learn event rules based on the analysis of trajectory series: trajectory segments were first clustered into primitive "events" (trajectory patterns), and then, a grammar induction algorithm produced a set of event rules. Makris and Ellis [85] considered spatial extensions of trajectories to construct path models, which were updated when new trajectories were matched. A similar approach is used in [90]. More recently, Basharat *et al.* [16] modeled a probability density function at every pixel location by means of GMM, considering not only spatial coordinates but also object sizes.

Other techniques have also been used to cluster trajectories and detect abnormality, especially Markov models. Porikli [105] used HMM to achieve event detection based on the unsupervised clustering of variable length trajectories, in a high-dimensional feature space. Piciarelli and Foresti [104] presented an online modeling algorithm to obtain a hierarchy of typical paths. Hu *et al.* [57] obtained motion patterns by spatially and temporally clustering trajectories using fuzzy K-means. Yao *et al.* [138] applied Markov models to learn contextual motion, in order to improve the results of low-level tracking and to detect abnormality. A detailed comparison of recent distance metrics and trajectory clustering techniques is available in [92], and Table 2.3 compiles the basic characteristics of the works discussed in the field of trajectory clustering and further activity interpretation.

Once a proper trajectory representation is chosen, most works focus on assigning semantic properties to the regions in which agent motion has been detected. Wang *et*

Trajectory-based activity recognition		
Main techniques	*Examples of recognized events*	*Reference to publication*
GMM	Typical paths, *anomaly*	Stauffer/Grimson [120], 2000
Self-organizing ANN	Typical paths, *anomaly*	Hu *et al.* [56], 2004
GMM	Typical paths, *anomaly*, *enter*, *exit*, *inactive*	McKenna/Charif [90], 2004
HMM	*Anomaly*	Porikli [105], 2004
GMM	Typical paths, *enter*, *exit*, *stop*	Makris/Ellis [85], 2005
Fuzzy K-means	Typical paths, *anomaly*,	Hu *et al.* [57], 2006
On-line clustering	Typical paths, *anomaly*,	Piciarelli/Foresti [104], 2006
GMM	*Car anomaly* (size, direction, speed)	Basharat *et al.* [16], 2008
Markov models	*Car moving off-road*, *car collision*, *traffic rule violation*	Yao *et al.* [138], 2008

Table 2.3

Representative sample of publications on trajectory-based activity recognition.

al. [132] proposed a method to automatically learn far-field semantic scene models by analyzing distributions of positions and directions of tracked objects on their trajectories, thus recognizing roads, walking paths, and sources/sinks. Li et al. [78] modeled activity correlation and abnormality by first segmenting the scene into event-based regions, modeled as a Mixture of Gaussians with features like aspect ratio and mean optic flow. Although the number of regions is learned automatically, their description is numerical. Similarly, the semantic region modeling in [131] is accomplished by clustering trajectories into different activities. Observed positions and directions are quantized, and the semantic regions are found as intersections of paths having similar observations. The analysis does not include temporal considerations.

Other works on region labeling tackle the detection of entry and exit zones. Makris *et al.*[85] used EM and GMM to cluster typical entry/exit areas and usual stopping zones. In [140], a combination of GMM and graph cuts are used to learn characteristic paths and entry/exit points for abnormality detection. Gryn *et al.* [46] used handcrafted direction maps to detect patterns such as using a particular door, or making an illegal left turn at an intersection. These direction maps were regularly spaced vector fields representing the direction of motion at locations of interest, and are scene specific, detecting motion in a particular image plane location.

All the aforementioned approaches build models able to segment vehicle or pedestrian paths, waiting zones, and entry/exit points. Nevertheless, they all disregard the inherent semantics that come associated to specific places and regions –e.g., chairs, crosswalks, bus stops–, which are not exploited by bottom-up analyses. Bottom-up techniques typically employ clustering to model the spatial distribution of single trajectories, thus making it possible to find common paths and detect abnormal occurrences; but their potential for behavior recognition is far from that of top-down, model-based approaches, which do exploit the region semantics, manually encoded by experts.

Pixel-wise scene segmentation		
Main techniques	*Examples of recognized objects or locations*	*Reference to publication*
MRF + PS	*Cows* vs. *horses*	Kumar *et al.* [70], 2004
LCRF	*Car*, *face*, *building*, *sky*, *tree*, *grass*	Winn/Shotton [135], 2006
Randomized forests	*Road*, *building*, *sidewalk*, *pedestrian*, *fence*, *bicyclist*	Brostow *et al.* [25], 2008
HoF + K-means + MRF	*Train platforms/railways*, *bus stops*, *park benches*	Dee *et al.* [30], 2008
Semantic textons	*Building*, *tree*, *sheep*, *bicycle*, *road*, *boat*	Shotton *et al.* [114], 2008
Version space + ontologies	*doors*, *balconies*, *stairs*, *canopies*, *railing*, *sign*	Hartz *et al.* [51], 2009
P^n potentials	*Sky*, *building*, *tree*, *grass*, *bird*	Kohli *et al.* [65], 2009
HoG + CRF	*building*, *car*, *road*, *sky*, *fence*, *pedestrian*, *cyclist*	Sturgess *et al.* [121], 2009

Table 2.4

Representative sample of publications on pixel-wise semantic segmentation. Most works in this field are based on appearance, few use dynamic information: we only found the last four in this list.

In order to exploit semantics inherent to locations, several authors have considered the problem of adding semantic characteristics to locations in a scenario. Towards robust pixel-wise segmentation and recognition of semantic regions, efficient techniques have been developed, such as MRF [69] and its variants, like DRF [70] or LCRF [135], or alternatives like TextonBoost[115] or Semantic Textons [114]. Improved techniques have been proposed, such as robust higher order potentials by Kohli *et al.* [65]. However, whilst there is a large literature aiming to semantically label multiple regions in images, it is difficult to find works that address this problem in videos, and using only dynamic information.

Dynamic data is incorporated by the following authors. Brostow *et al.* [25] complemented appearance-based features with their motion and structure cues to improve object recognition. Dee *et al.* [30] worked on unsupervised learning of semantically meaningful spatial regions –e.g., *train platforms*, *bus stops*– in videos, only from motion patterns. These patterns were quantized within the cells of a scene grid. Hartz *et al.* [51] investigated automatic learning of complex scenes with structured objects like *doors*, *balconies*, *stairs*, or *canopies* using ontological constraints, for images and using only appearance. Sturgess *et al.* [121] presented a framework for pixel-wise object segmentation of road scenes, combining both motion and appearance features. They partitioned monocular image sequences taken from a car into regions like *building*, *car*, *road*, *sky*, *fence*, etc. using CRFs. Table 2.4 summarizes a representative selection of publications in this field.

Our proposal contributes to the field of urban surveillance by building, automatically and in a fully unsupervised manner, semantic scene models uniquely based on dynamic information from trajectories. The resulting models are richer than simple source–path–sink ones. In this paper we show a novel technique that, having learned the spatiotemporal patterns of moving objects, infers the semantic meaning of background regions, such as *pedestrian crossings*, *sidewalks*, *roads* or *parking areas*; this process is guided by a taxonomy to incorporate the semantic properties to be reported. In our case, the categorization of regions from their statistical models is posed as a labeling task and formulated as a MAP-MRF inference problem, defined by irregular sites and discrete labels [79].

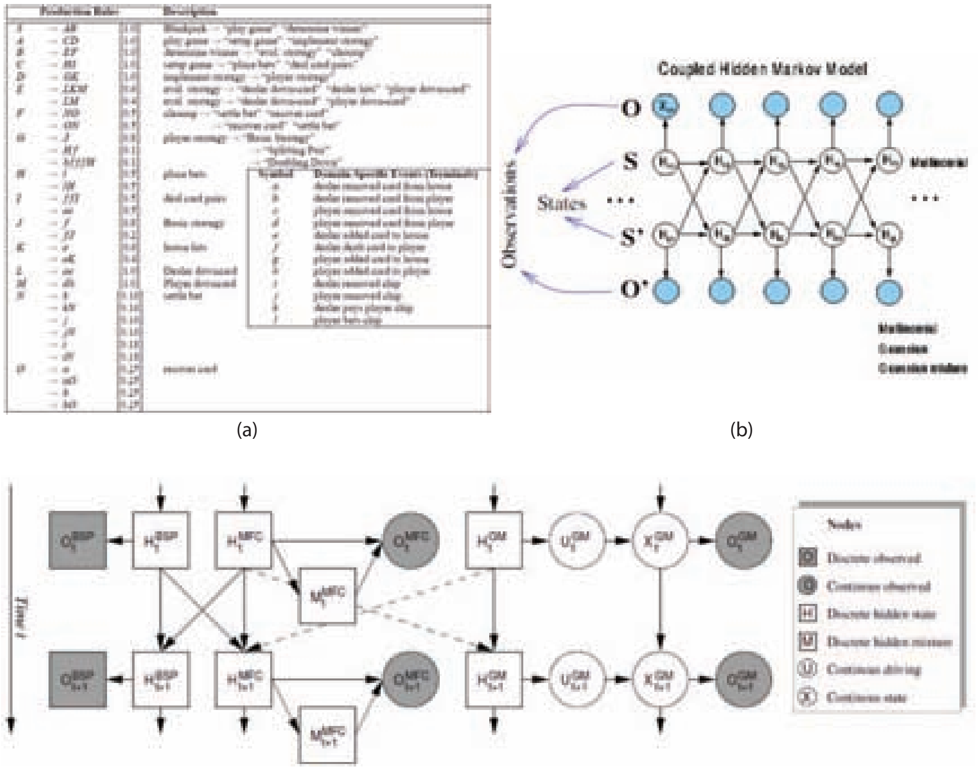
2.3 Modeling high-level events

Algorithms for detection and tracking have been greatly improved during the last years, and although there are still many issues to cope with –e.g., appearance variability, long-term occlusions, high-dimensional motion, crowded scenes–, robust solutions have been already provided that capture the motion properties of the objects in dynamic and complex environments [107, 108]. But to understand scenes involving humans, to interpret "*what is happening in the scene*", we need more abstract and meaningful schemes than purely physical laws. To understand long image sequences showing semantic developments, we require another abstraction scheme: the *event* [103]. An event is regarded as a conceptual description summarizing the contents of a development, and that description is closely related to real world knowledge.

The recognition of events in video sequences has been extensively tackled by the research community, ranging from simple actions like walking or running [97] to complex, long-term, multi-agent events [75]. The recognition of complex events and behaviors is becoming more and more a hot topic of the literature in this field. Three main approaches are generally followed towards the recognition of non-basic events: pattern recognition methods, state models, and semantic models.

First of all, the modeling formalisms used include many diverse techniques for pattern recognition and classification, such as neural networks and self-organizing maps [143], K-nearest neighbors (kNN) [88], boosting[118], support vector machines (SVN) [97], or probabilistic or stochastic context-free grammars (CFG) [64, 91]. In addition, the statistical modeling of Markov processes is tackled using state models, such as hidden Markov Models (HMM) [100, 136], Bayesian networks (BN) [52], or dynamic Bayesian networks (DBN) [3] have been often used when pursuing the recognition of actions and activities. All these have been successfully applied to the domain of event recognition, as it can be seen in Table 2.5.

Nevertheless, the high complexity found in the domain of video sequences stresses the need to employ richer –in the sense of *more explicit*– semantic models. This need comes emphasized by the fact that the interpretation of activities depends strongly on the locations where events occur –e.g., traffic scenes, airports, banks, or border controls in the case of surveillance–, which can be efficiently exploited by means of conceptual models. Therefore, it is reasonable to make use of domain knowledge in order to deal with uncertainty and evaluate context-specific behaviors. Thus, a series of tools based on symbolic approaches have been proposed to define the domain of events appearing in selected environments, e.g. those based on conceptual graphs or conditional networks.



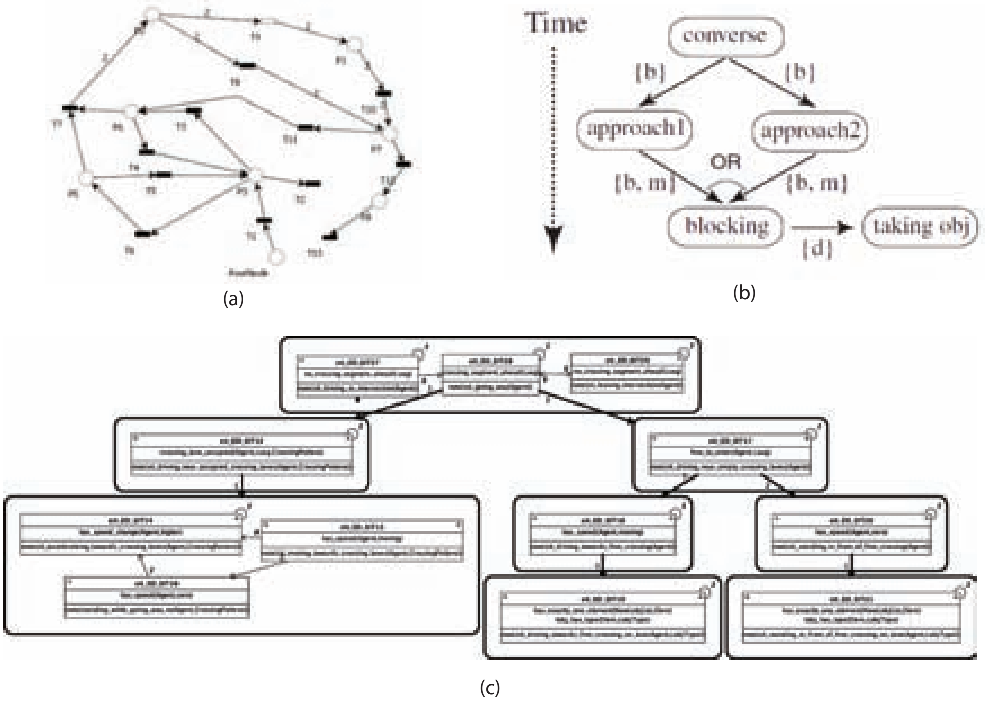
(a) (b)

(c)

Figure 2.1: Probabilistic techniques for event recognition. (a) Events in a Blackjack play modeled via SCFG [91]. (b) Coupled HMM to detect interaction events between individual humans. [100]. (c) State diagram of a DBN to recognize *meeting* events [3].

Starting from the early use of Finite State Automatons [58, 52] and similar improved symbolic graphs [83], researchers have increased the expressivity of the models, so that they can manifest precise spatial, temporal, and logical constraints. Such constraints have ended up complementing each other in multivariate analyses, e.g., by means of temporal constraint satisfaction solvers applied over symbolic networks [130]. More recently, Nagel and Gerber [95] proposed a framework that combines Situation Graph Trees (SGTs) with Fuzzy Metric Temporal horn Logic (FMTL) reasoning, in order to generate descriptions of observed occurrences in traffic scenarios. Extensions of Petri Nets have also been a common approach to model multi-agent interactions, and used as well for human activity detection [4]. Some other recent approaches have employed symbolic networks combined with rule-based temporal constraints, e.g. for activity monitoring applications [40]. Fig. 2.2 shows examples of these symbolic structures used for the automatic recognition of events.

All these symbolic models, which work with predefined behaviors, show good performances at behavior recognition, provide explanations of the decisions taken, and allow uncertainty to be incorporated to the analysis, thus making it more robust to



(a)

(b)

(c)

Figure 2.2: Model-based techniques for event recognition. Petri Net modeling a security check at the entrance of a building [74]. Graphical representation of a multi-thread event *stealing* [52]. SGT specializing a `sit_giving_way` situation [94].

noisy or incomplete observations. The reasoning and interpretation modules conceived in this thesis follow the work done by Nagel, and its posterior adaptation to human behaviors accomplished by Gonzàlez's HSE [43]. They integrate fuzzy logic inference engines with SGT to model the semantics of different events. We choose SGTs over other symbolic approaches due to the efficacious mechanisms of specialization and prediction they incorporate, which help modeling the universe of situations in a clear, flexible, and controllable manner. SGTs and fuzzy metric-temporal logic, unlike Petri nets, are adapted to model and evaluate human behaviors on specific contexts, which we provide by means of ontologies.

The cited symbolic approaches allow semantic representations of the events detected, which facilitate implementing user-computer interfaces. Nonetheless, none of them carries out a thorough evaluation of the correctness or suitability of the selection of events, mainly due to the limited amount of semantics found in the video sequences. Other works have proposed lists of semantic events for the surveillance domain directly proposed by specific groups [128], or based on the system capabilities to generate them [111, 38]. We propose instead to base the models on evidence provided by human participants.

Recent, relevant work dealing with concept selection is presented in [68], which comprises approaches included in two acknowledged EU projects, aceMedia and MESH,

Video event recognition		
Main topic	*Examples of recognized events*	*Reference to publication*
BN + FSM	*Converse*, *steal*, *approach*, *take object*	Hongeng/Nevatia [52], 2001
Stochastic CFG	*Player added card*, *dealer removed chip*, *player bets chip*	Moore/Essa [91], 2002
kNN	*Run*, *skip*, *march*, *hop*, *side-walk*	Masoud *et al.* [88], 2003
Symbolic networks	*Attack*, *robber enters*, *cashier at safe*	Vu *et al.* [130], 2003
Multi-layered FSM	*Walking*, *parking*, *theft*	Mahajan *et al.* [83], 2004
Conditional MRF	*Bend pick side*, *dancing*, *jump forward*, *side walk*	Sminchisescu *et al.* [117], 2006
Boosting	*Talk on phone*, *scratch*, *take medication*, *yawn*, *put eyeglasses*	Smith *et al.* [118], 2005
DBN	*Put down*, *press button*, *pick up*	Vincze *et al.* [129], 2006
DML–HMM	*Can taken*, *browsing and paying*, *moving cargo lift*, *truck comes*	Xiang/Gong [136], 2006
Transition graph	*Corner kick*, *golf swing*, *excited speech*	Xiong *et al.* [137], 2006
Symbolic networks	*Arrive*, *enter area*, *manipulate container*, *stop*	Fusier *et al.* [40], 2007
FSM + SVM	*Crouch*, *wave*, *pick up*, *reach*	İkizler/Forsyth [58], 2007
DBN	*Crack egg2*, *pour milk*, *stir*, *flip bread2*, *pickup vanilla*	Laxton *et al.* [75], 2007
Petri nets	*Customer / bank employee interaction*, *bank robbery attempt*, *bank robbery success*, *access safe*	Albanese *et al.* [4], 2008
SGT	*Wait*, *cross*, *leave*, *sit down*, *walk among chairs*	Gonzàlez *et al.* [43], 2008
Stochastic CFG	*Money found in tray*, *remove money*, *take receipt*, *pick up scanner*	[64], 2008
SGT	*Change lange*, *turn*, *catch up*, *follow*, *lose a lead on*	Nagel/Gerber [95], 2008
pLSA + SVM	*Walking*, *boxing*, *hand clapping*, *camel-spin*, *sit-spin*	Niebles *et al.* [97], 2008
Self-organizing maps	*Washing dishes*, *toileting*, *preparing a snack*, *doing laundry*, *lawnwork*	Zheng *et al.* [143], 2008
Petri nets	*Enter*, *long security check*	Lavee *et al.* [74], 2009
SVM	Shot on goal, placed kick, throw in, goal kick, protest, airplane flying, running	Bertini *et al.* [15], 2010

Table 2.5

Representative sample of publications focused on the recognition of activities, events, and behaviors.

both dealing with semantic video retrieval. Diverse suggestions are given regarding the type and frequency of desirable concepts to include in semantic models. Some of their main results have been used to justify the organization of knowledge in our approach. Our contribution to the field of concept selection for model building is presented in Section 4: we propose a pipeline that not only exploits the semantics of textual descriptions from human participants, but additionally guides experts while defining and integrating rule elements in behavioral models.

The ontological cognitive vision system presented in this thesis builds on both purposive and reactive data flows, which incorporate techniques from several vision and reasoning levels. Most authors agree that mechanisms for the evaluation, gathering, integration and active selection of these techniques are fundamental to attain robust interpretation of dynamic information [129, 45]. These needs for coordination of contextual knowledge suggests to single out specific stages for semantic manipulation. Although many advanced surveillance systems have adopted semantic-based approaches to face high-level issues related to abstraction and reasoning, the use of ontologies at high levels of such systems is only now beginning to be adopted. Following these premises, the structure of the proposed system is based on a modular architecture, which allows both top-down and bottom-up flows of information, and has been designed to integrate ontological resources for cooperation with the reasoning stage.

2.4 Interacting with end-users

A prediction for the future that is widely accepted today is that the computing techniques as we know them will move progressively to the background, while special attention will be drawn on the human user. [101]. What this prediction suggests is that next generation of computing to come will be especially focusing natural means of interaction with end-users, using interfaces that are based on human models and pursue human-oriented communication. In this context, the use of natural language and virtual and interactive environments is vital to achieve that goal.

Next we list a brief selection of works pursuing an advanced interaction with end-users in fields related to video understanding, video surveillance, multimedia, and their derived applications. They include systems for automatic generation of textual information, dialogue systems, augmented reality, or virtual storytelling for simulation, for instance.

Natural language extensions

The automatic analysis and description of temporal events was already tackled by Marburger et al. [86], who proposed a NL dialogue system in German to retrieve information about traffic scenes. Other early publications like [54] describe work on discourse generation using discourse structure relations, especially regarding automated planning and generation of text containing multiple sentences. More recent methods for describing human activities from video images have been reported by Kojima et al. [66]; [80] discusses a general framework for semantic interpretation of vehicle and pedestrian behaviors in visual traffic surveillance scenes; and a series of automatic visual surveillance systems for traffic applications have been studied in [5] and [27], among others. These approaches present one or more specific limitations such as textual generation in a single language, surveillance for vehicular traffic applications only, restrictions for uncertain data, or very rigid environments.

There have also been intense discussions about how to interrelate the semantic information extracted from video sequences. The aceMedia integrated project intends to unify multimedia representations by applying ontology-based discourse structure and analysis to multimedia resources [67]. The EU Project ActIPret uses semantic-driven techniques to automatically describe and record activities of people handling tools in NL, by exploiting contextual information towards symbolic interpretation of spatiotemporal data [129]. Its reasoning engine focuses on the coordination of visual processes to obtain generic primitives from contextual control. The intelligent multimedia storytelling system CONFUCIUS interprets NL inputs and automatically generates 3D animation and speech [81]. Several methods for categorizing eventive verbs are discussed, and the notion of visual valency is introduced as a semantic modeling tool.

In [103], Park and Aggarwal discuss a method to represent two-person interactions at a semantic level, also involving user-friendly NL descriptions. Human interactions are represented in terms of cause-effect (event) semantics between syntactical agent–motion–target triplets. The final mapping into verb phrases is based on simultaneous and sequential recognitions of predefined interactions. Concerning the semantic mappings of NL sentences, it is also interesting to mention Project FrameNet [11] and its successor, WordNet [34], which has built a lexical resource for several specific languages such as English, Spanish, German, or Korean, aiming to list the acceptable semantic and syntactic valences of each word in each of its contexts. The automatic exploitation of this repository for applications involving visual data has been done before, for instance in Hoogs *et al.* [53], who tackled the translation of visual information into words using WordNet. The resulting words are used to generate scene descriptions, by searching through the semantic relationships in this repository.

Virtual environments

The synthetic generation of virtual environments is also significant in the field of user interaction, providing tools for visual communication or simulation, for instance. Following [82], some of the most clear future challenges in creating realistic and believable Virtual Humans consist of generating on-the-fly flexible motion and providing them with complex behaviors inside their environments, as well as making them interactive with other agents. Interaction between real and virtual agents has been little considered previously [41, 13]. Gelenbe et al. [41] proposed an augmented reality system combining computer vision with behavior-based agents. Behavior is modeled using a hierarchy of three behavior modules, but without considering the particular features of human motion and behavior. Zhang et al. [139] presented a method to merge virtual objects into video sequences recorded with a freely moving camera. The method is consistent regarding illumination and shadows, but it does not tackle occlusions with real moving agents. The use of computer vision techniques in augmented reality has also been recently confronted by Douze et al. [32], where moving targets are tracked from image sequences and merged into other real or virtual environments. However, the method does not consider behavioral virtual agents in the resulting sequence.

Some research has also been done on combining approaches from augmented reality and virtual storytelling technologies. Balcisoy et al. [12, 13] present augmented reality frameworks in which external users restrict virtual agents to perform a given script, converting them in directors of the scene. Papagiannakis et al. [102] mixes the two approaches to present virtual actors that introduce visitors of ancient locations into the world of fresco paintings, by providing these actors with dramaturgical behaviors. Lee et al. [77] describe a Responsive Multimedia System for virtual storytelling, in which external users interact with the system by means of tangible, haptic and vision-based interfaces.

2.5 Ontologies to enhance video understanding

It has been repeatedly stated how ontologies can be used effectively for relating semantic descriptors to image or video content, or at least use them to represent and fuse structured prior information from different sources towards that end [123]. Several classical methods from artificial intelligence to represent or match ontological knowledge –e.g., Description Logics (DL), frame-based representations, semantic networks– are becoming popular again since the start of the *Semantic Web* initiative [68, 9]. Nevertheless, the challenge today is how to apply these approaches to highly ambiguous or uncertain information, like that coming from language and vision, respectively. For this reason, the incorporation of ontologies into cognitive vision systems has also awaken the interests of many researchers in the field [84, 119]. The use of DL to model uncertainty has been long discussed; an overview of the research in this field is presented in Baader *et al.* [8].

In the case of video surveillance, ontologies have been used to assist to the recognition of video events. Several authors have engaged initiatives to standardize taxonomies of video events, e.g., [96] proposed a formal language to describe event ontologies, VERL, and a markup language, VEML, to annotate instances of ontological events. The use of this language is exemplified in videos from the security and meeting domains. Ma and McKevitt [81] present an ontology of eventive verbs for multimodal storytelling system including visual and linguistic concepts.

Regarding the field of multimedia, the automatic processing of multimedia content has been enhanced by the apparition of new multimedia standards, such as MPEG-7, which provide basic functionalities in order to manipulate and transmit objects and metadata, and measure similarity in images or video based on visual criteria. However, most of the semantic content of video data is out of the scope of these standards. In these cases, ontologies are often used to extend standardized multimedia annotation by means of concept hierarchies [124, 59], and also to provide meaningful query languages –e.g., RDQL or SWRL– as tools to build, annotate, integrate, and learn ontological information. An overview of such languages is presented in [142].

There have been efforts towards the generation of textual representations and summaries from ontologies [22, 133]. In fact, these approaches are general-purpose ontology verbalizers, agnostic of the class types and their properties, which result in outputs that are in general too verbose and redundant. Our contribution adapts the textual descriptions and summaries to the type of content described, regarding its organization into the modeled domain ontology.

Ontology-based approaches are also suitable for designing processes to query, report, or mine data from distributed and heterogeneous sources. These capabilities derive a series of tasks that are usually requested in the domain of multimedia semantics, such as automatic video annotation to enable query-based video retrieval. Bertini et al. [17] have recently presented an ontology-based framework for semantic video annotation based on the learning of spatio-temporal rules. First Order Inductive Learner (FOIL) is adapted to learn rule patterns that have been then validated on some TRECVID video events. Similarly, other approaches emphasize the use of ontologies to enable forensic applications in video surveillance [126].

The understanding of linguistic events has also been approached with ontologies. For instance, Cimiano *et al.* [28] presented an ontology-driven approach that, based on Discourse Representation Theory from linguistics, computes conceptual relations between events extracted form a text and a referring expression representing some other event, a state or an entity. Recent large-scale endeavors like the Virtual Human Project [49] propose a complete architecture for virtual humans, including NL capabilities for generation and understanding, speech recognition and text-to-speech synthesis, task reasoning, behavior blending, and virtual environment generation. An ontological design was chosen for flexibility and extensibility, and to deal with the many multimodal representations of knowledge considered. This work stresses the importance of ontologies especially when relating language and concepts.

Resum

En l'actualitat podem trobar un gran nombre de publicacions en matèria d'anàlisi de contingut semàntic en seqüències de vídeo, sobretot pel que fa a aplicacions destinades als camps de la video vigilància i el processament multimèdia. En general, els mètodes d'anàlisi computacional es poden classificar en dos grans grups, segons es basin en models predefinits o bé recorrin a tècniques probabilístiques. En aquest capítol es recull breument part de la feina investigadora més rellevant en els dos casos per l'objectiu descrit. Tot i això, el contingut d'aquesta tesi es basa majoritàriament en l'ús de models semàntics predefinits, que malgrat requerir la construcció prèvia dels models per part d'experts, possibilita descripcions semàntiques més complexes i expressives i fa que l'entorn sigui molt més controlable. Això ens resulta de gran ajuda especialment en el cas de la video vigilància.

L'extensa recerca existent en el camp s'ha dedicat a solucionar tot un seguit de problemes de caire divers, com ara el reconeixement d'activitats de variada complexitat, el context semàntic de les zones observades a partir de característiques de moviment (necessari per a optimitzar l'anterior tasca), o els mecanismes de comunicació efectiva amb usuaris finals, per mitjà de tècniques com el llenguatge natural o la realitat virtual o augmentada. En el cas del reconeixement de regions semàntiques d'interès, generalment es realitza un pas de clusterització (K-means, GMM, xarxes neuronals) i un pas posterior de segmentació, generalment basat en tècniques de

Markov (MRF, CRF) o arbres randomitzats de característiques d'imatge.

Quant a l'avaluació d'activitat humana, durant la darrera dècada s'ha aplicat una llista pràcticament interminable de tècniques de tot tipus: cadenes de Markov, xarxes bayesianes, pàrsing probabilístic, boosting, xarxes neuronals, xarxes de Petri, SVM, arbres de situació. . . Per al nostre conjunt d'aplicacions, generalment de domini tancat i restringit, la nostra preferència passa per fer servir models simbòlics.

Un altre aspecte important té a veure amb l'organització de la informació, que ha de facilitar el manteniment i l'extensibilitat dels models, especialment en el cas de sistemes multi modals com el nostre. En aquestes circumstàncies, l'ús d'ontologies resulta adequat per a una correcta centralització i reutilització de la informació disponible, tal i com s'ha demostrat repetidament amb la incorporació d'interfícies lingüístiques i mòduls d'explotació semàntica per a sistemes cognitius artificials de visió.

Chapter 3

Taxonomy-based dynamic semantic region learning

"A place for everything, and everything in its place."

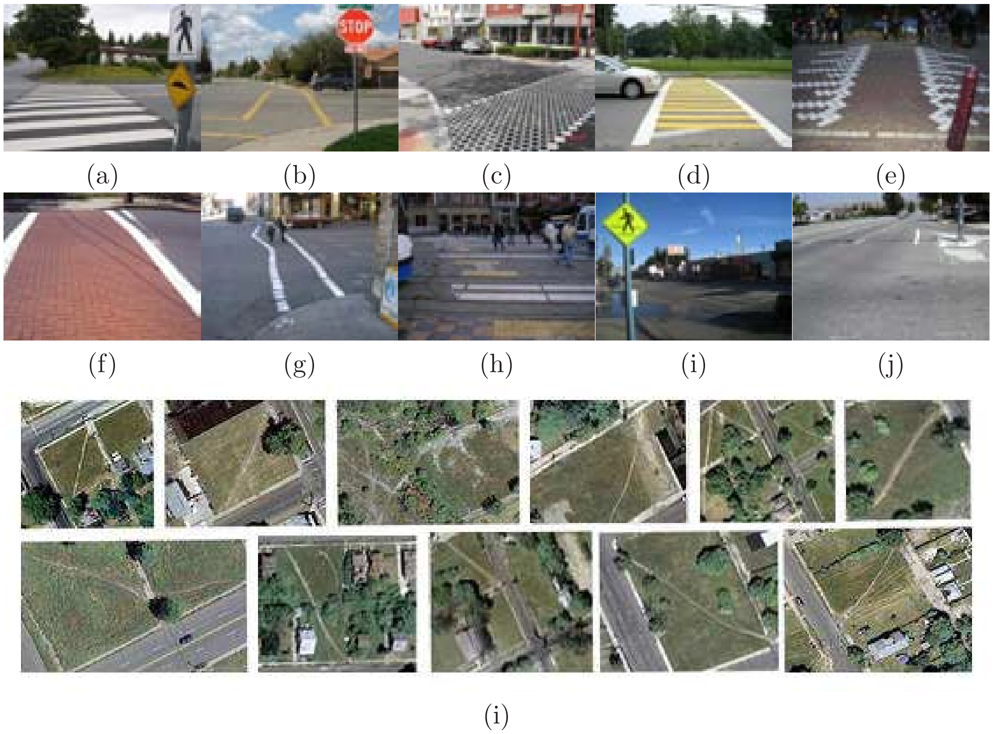
Thrift (1875), by Samuel Smiles

Systems for advanced activity recognition depend strongly on the particular configuration of a scenario. In such systems, the locations where interesting motion events occur are located and attributed with semantic properties by human experts. This chapter explores the automation of this process, i.e., segmenting and labeling semantic regions in scenarios from a given domain –urban surveillance– using only common knowledge to guide the analysis of image sequences. Hence, both the sensory *and the* semantic *gap intervene in this chapter –i.e., interpreting the semantics of a region from limited observations of motion–, along with the* model *gap, which limits a priori knowledge to the domain of urban traffic.*

As stated in [30], the ability to reason about what we see in video streams is influenced by our ability to break down spatial structures into semantically meaningful regions. Such regions are characterized by their appearance, e.g., the line markings of a crosswalk allow us to identify it, visually. Nevertheless, we can also identify regions functionally, i.e., according to the behavior observed on them. This is clearly the case for crosswalks, see Fig. 3.1 (a–j).

Urbanism nowadays is packed with examples of how the observed behaviors motivate the functionality of a region, especially speaking of paths. For instance, the paths designed in Dartmouth University campus were placed according to the grounds left bare by their walking students in winter. In addition, those well-worn paths that develop when people depart from formal routes and create their own, unofficial, more straight paths, were called by Gaston Bachelard *chemins du désir* or *pathways of desire*, see Fig 3.1 (i).

In this chapter, we propose to exploit observed behaviors performed by pedestrians and vehicles in urban scenarios, in order to recognize and label meaningful regions in them. An automatic modeling of semantic regions in the scenario is beneficial



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(i)

Figure 3.1: (a–j) Different instances of crosswalks in urban scenarios around the world. Although their appearance varies significantly (a–g) or they are not clearly visible (h–j), their functionality stays the same: cars should stop when pedestrians cross. (i) *Pathways of desire* in Detroit.

for posterior reasoning systems, which facilitate knowledge-based interpretations of complex occurrences in a situated context.

3.1 Background labeling by compatibility

The semantic learning of a background model consists of partitioning an arbitrary scenario of the domain into a lattice of regions, and have each region learn a spatiotemporal model. Each model should be estimated based on trajectory properties, and finally assigned an explicit label that categorizes it. Here, we tackle the problem of *semantic region learning* as one of *multiclass semantic segmentation*. Towards this end, efficient techniques have been developed, such as MRF [69] and its variants, like DRF [70], or LCRF [135], or alternatives like Semantic Textons [114]. In our case, the categorization of regions from their statistical models will be posed as a labeling task and formulated as a MAP-MRF inference problem, defined by irregular sites and discrete labels [79].

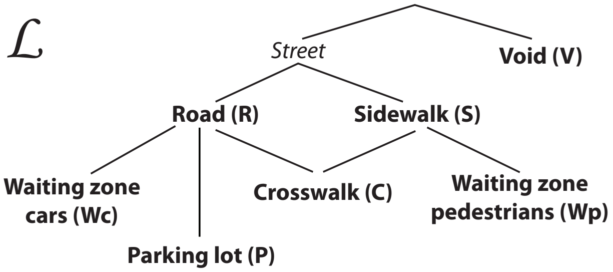


Figure 3.2: Taxonomy of location categories for urban surveillance.

Sites and labels

The lattice of irregular regions to be labeled is usually defined either by perceptual groups –out of a segmentation process–, or by clusters of recognized features within the scene [79]. Instead, we aim to define lattices that capture the condition of far-field projectivity, which is characteristic of scenarios in our domain. To do so, we compute the scene to ground-plane homography [50], so that each lattice is a set of regions $\mathcal{R}$ obtained as the projection of a rectangular grid from ground-plane to scene.

In addition to the sites, a set $\mathcal{L}$ of seven discrete labels defines *generic*, *common*, and *relevant* locations in urban surveillance. Labels are organized taxonomically as shown in Fig. 3.2. A void label (V) is made available for those cases in which none of the labels applies, as in [35].

Inference

Having defined the set of sites and labels, we next describe the process of assigning a label $l \in \mathcal{L}$ to each region $r \in \mathcal{R}$. The disparity of labels is assumed to be piecewise smooth in the lattice of regions. A series of observation vectors $o = \{x, y, a\}$ constitutes the evidence from the trajectories, where (x, y) is the estimated position of the agents in the image plane –the lower middle point of their bounding box–, and a is a binary parameter stating whether the agent is a vehicle or a pedestrian. The derivation of the site labels $\{l\}$ is formulated as a MAP-MRF inference in terms of a pairwise Markov network, whose graph configuration is factored into the joint probability

$$P(\{l\}, \{o\}) = \frac{1}{Z} \prod_{r \in \mathcal{R}} \phi_r(l_r, o_r) \prod_{\{r,s\} \in \mathcal{N}} \psi_{r,s}(l_r, l_s), \tag{3.1}$$

where Z is a normalization factor. The *data compatibility* function $\phi_r(l_r, o_r)$ is interpreted as the likelihood of choosing label l for region r given the vectors o observed in r. This function is learned by trajectory analysis, as later explained in Section 3.2.

On the other hand, smoothness constraints are encoded into $\psi_{r,s}(l_r, l_s)$, so-called *internal binding*, which models how neighboring regions affect to each other regarding their classes. In this term, the set $\mathcal{N}$ contains all pairs of interacting regions, in our case adjacent 8–connected regions in the projected grids. In our work, $\psi_{r,s}(\cdot)$ is a prior set of constraints directly taken from topological assumptions. These are derived

Code and label from $\mathcal{L}$			pedestrian	vehicle	stop	parking	
C	Crosswalk	$f^1 = [$	+	+	+/−	−	]
S	Sidewalk	$f^2 = [$	+	−	−	−	]
R	Road	$f^3 = [$	−	+	−	−	]
W_p	Ped. waiting zone	$f^4 = [$	+	−	+	−	]
W_c	Veh. waiting zone	$f^5 = [$	−	+	+	−	]
P	Parking	$f^6 = [$	+/−	+	+/−	+	]
V	Void	$f^7 = [$	−	−	−	−	]

Table 3.1

Description of labels as prototypical vectors of the trinary features *pedestrian*, *vehicle*, *stop*, and *parking*.

from a defined hierarchy of labels depicting domain knowledge, as later explained in Section 3.3.

Once the compatibility functions $\phi_r(\cdot)$ and $\psi_{r,s}(\cdot)$ are defined, a max-product belief propagation (BP) algorithm [35] derives an approximate MAP labeling for Eq. (3.1).

3.2 Data compatibility

We define the function $\phi_r(l_r, o_r)$ as the likelihood of region r to be labeled as l, having observed a series of vectors o_r in the region, and according to a motion-based model that encodes prior domain knowledge.

Challenges arisen by semantic scene –similarly, by document analysis or medical imaging– deal with classes that are overlapping and not mutually exclusive. Hence, we characterize scenario regions following the prototype theory, in which class labels are defined in terms of conjunctions of required (+), forbidden (–), and irrelevant (+/−) features [29]. In our case, labels are modeled using 4 features: target is (i) a *pedestrian* or (ii) a *vehicle*, (iii) has *stopped*, and (iv) has *parked*, i.e., has stopped longer than a predefined time value, see Table 3.1. A series of prototypical feature vectors $\{f^1 \dots f^{|\mathcal{L}|}\}$ results from this step.

Next step consists of online smoothing and sampling data retrieved from tracking. To do so, each new complete trajectory is fitted by iteratively increasing a sequence of connected cubic b-splines (Fig. 3.3b): an adjustment step divides a spline into connected sub-splines more fitted to the trajectory, and a termination step validates a subsequence when its maximum distance to the trajectory is below a 10% of the total length. Once the recursion is done, the global sequence of splines is sampled to generate a set of time-equidistant control points (Fig 3.3c), each one having an observation $o = (x, y, a)$. The position (x, y) is estimated by a multi-target tracker [107], and the target type (a) is identified using a scene-invariant discriminative approach [23].

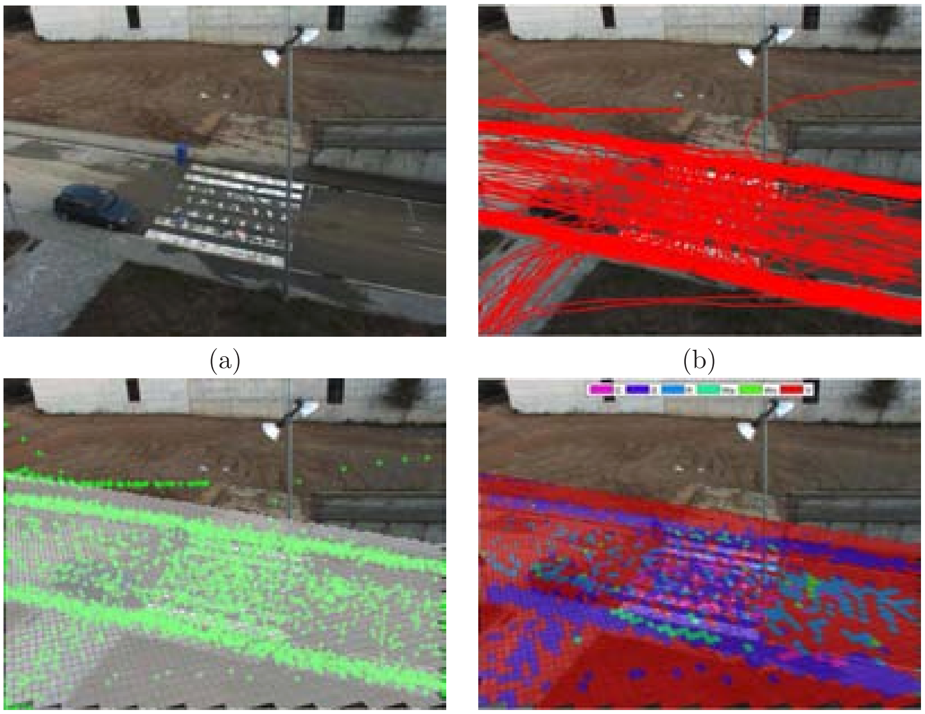

(a) (b)
(c) (d)

Figure 3.3: Region modeling by trajectory analysis: (a) original image, (b) smoothed trajectories, (c) sampled control points, (d) initial labeling.

When a new control point is generated, its enclosing region updates an histogram of the 4 features described. The two last ones are derived from consecutive observations: a *stop* property is asserted when a position is repeated, whereas a *parking* is told to happen when a target is stopped for more than 2 minutes. Finally, an online averaged vector of observed features f_o is obtained for each region.

The data compatibility of the observations in region r with label $l \in \mathcal{L}$ is a softmax function of the Hamming distance between the averaged vector of features observed, f_o, and the vector defined for that label, f^l:

$$\phi_r(l_r, o_r) = \frac{\exp(-d_H(f_o, f^l))}{\sum_{m \in \mathcal{L}} \exp(-d_H(f_o, f^m))}. \tag{3.2}$$

The data compatibilities learned are used to initially provide a rough scene model. This initial labeling omits the inference phase, and simply assigns to each region the label with a highest value of $\phi_r(\cdot)$, see Fig. 3.3d. Due to the limited coverage of the scene by the control points, there is a massive presence of *Void* labels, in red.

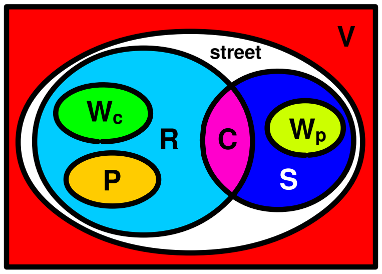


Figure 3.4: Topological constraints equivalent to the taxonomy of labels in Fig. 3.2.

3.3 Smoothness

The smoothness term $\psi_{r,s}(l_r, l_s)$ specifies inter-region compatibilities, stating how the system privileges or disfavors label l_r at expenses of l_s when r and s are adjacent. In other words, it conditions *a priori* the apparition of neighborhoods formed by a certain combination of classes. The goal here is to specify compatibilities that discard unlikely labelings, smooth poorly sampled ones, and preserve detailed information that be scarce but consistent.

In our case, advantage is taken on the hierarchical organization of $\mathcal{L}$ to constrain discontinuities between labels. $\mathcal{L}$ fixes topological constraints of set inclusion, as seen in Fig. 3.4: it establishes relations of particularization; e.g., a *parking lot* is a concrete segment of *road*, and also constrains the adjacency between different regions. Consequently, compatibilities are fully specified by

$$\psi_{r,s}(l_r, l_s) = \begin{cases} 1 & l_r = l_s \\ \alpha & Adj(l_r, l_s) \\ \beta & \text{otherwise} \end{cases} \tag{3.3}$$

where $1 > \alpha > \beta > 0$, and $Adj(l_r, l_s)$ states that l_r and l_s are adjacent in the topological map, i.e., direct links in the taxonomy. For example, P–R, C–R, or C–S are adjacent pairs, but W_c–P or R–S are not. This model tends to firstly maintain the identity of the sampled labels, secondly favor dilation and erosion between adjacent regions, and ultimately allow relabeling for region smoothness.

3.4 Geodesic interpolation

Having defined compatibilities for observed evidence and sought smoothness, the application of an efficient BP algorithm [35] approximates an optimal labeling via MAP-MRF inference. Nonetheless, certain issues make it difficult to obtain accurate segmentations. A stage is proposed before the inference step to overcome these difficulties.

In cases of very poor sampling, e.g., when estimating models of parking lots, the regions obtained by MAP-MRF inference with the smoothness prior are often still disconnected or not representative. To solve this problem, a preprocessing stage is

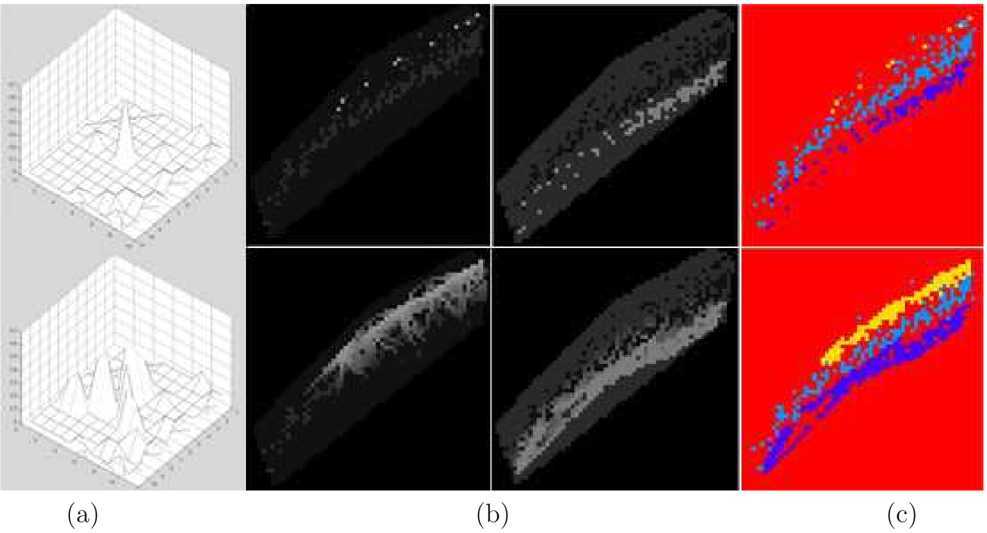


Figure 3.5: Top: non-smoothed marginal probabilities viewed (a) as a discrete mesh and (b) as intensity maps, and (c) initial label assignment (best viewed in color). Bottom: effects of the interpolation.

used to reinforce spatial coherence by geodesic interpolating lines; the idea is to create linear ridges that connect high-valued and isolated samples in each label's marginal probability map (Fig. 3.5a), in order to emphasize the presence of connected structures in them (Fig. 3.5b). As a result, the subsequent MAP-MRF process is reinforced with these structures and guides more sensible inferences for an eventual labeling, as shown in Fig. 3.5c.

3.5 Evaluation

The presented framework has been evaluated in 5 urban datasets, obtained both from private and public (web) cameras, and having diverse characteristics. The *Hermes* dataset [1] presents an interurban crosswalk scenario with more pedestrians than vehicles; *Oxford centre* [2] shows an intersection highly populated by both target types; *Devil's Lake* [3] presents moderate agitation but challenges with an intense projectivity; *Kingston–1* contains a partially seen bus stop close to a crosswalk, and *Kingston–2* shows a minor street with perpendicular parking lots used for long periods of time. These two last scenarios are extracted from the Kingston dataset [19]. Night sequences have been omitted.

Evaluation is carried out using 25 ground truth images, i.e., 5 participants per scenario, consisting of pixel-level maps segmented in the 7 categories shown in Fig 3.2. Participants were asked to visually identify the semantic regions by observing recorded

[1] http://www.hermes-project.eu/
[2] http://webcam.oii.ox.ac.uk/
[3] http://www.opentopia.com/showcam.php?camid=4182

footage, and partition them accordingly. In order to evaluate discriminant capability, and given that manual region labeling of the scenarios is prone to vary across humans, the system will perform well if segmentation errors compare to inter-observer variability. This criterion is commonly used for validation in biometrics [72]. To accomplish this, each ground truth image has been divided into the cells of its corresponding grid, and a modal filter has been applied over each cell, assigning the most repeated pixel label to that region. Finally, each label assignment has been evaluated against the other ground truths and averaged, for each ground truth and scenario. Fig. 3.6a shows the results of this inter-observer evaluation, which constitute a baseline of the desired system's performance.

The performance of our method has also been compared against a median filter. To do so, we have computed 3 different accuracy scores over the 5 datasets, evaluating both techniques against the ground truth assignments. In the evaluation tests, the maximum number of iterations for both the MAP-MRF and the median filter has been limited to 15. The values of α and β for the MAP-MRF are 0.80 and 0.60 respectively, in all the experiments.

The matricial configuration of the lattice reduces computational effort in both region modeling and label inference. Observations update the region models online as trajectories are complete. Regarding the final inference over regions learned, for a grid size of 75×75 geodesic interpolation takes at most 3 seconds to complete, and the BP algorithm with maximum iterations takes approximately 90 seconds in a Pentium II 3GHz machine with 2Gb RAM.

We analyze the consistency of the results by testing over a wide range of grid size values, which is the main parameter intervening in the sampling process: given that each control point sampled from a trajectory affects uniquely its enclosing region, the number of cells tesselating the scenario is indicative of the area of influence of tracked objects during region modeling. The dimensions of the projected grid in our experiments range from 40×40 to 150×150. Lower cell resolutions do not capture the details of the scenario, thus not being suitable to model semantic regions.

Additionally, the tracked trajectories used as observations incorporate an amount of tracking errors. Each error consists of one or more of the following cases: misclassification of agents, lost tracks, and false detections. Table 3.2 gives numerical information on the agents involved in each scenario and the number and type of erroneous observations. The system has been evaluated with and without the presence of errors, in order to test robustness.

Quality scores

The performance of each scenario has been evaluated in terms of accuracy. Three scores have been considered: overall accuracy (OA), segmentation accuracy (SA), and weighted segmentation accuracy (WSA). The two former scores are defined by

$$OA = \frac{TP+TN}{TP+FP+TN+FN}, \qquad SA = \frac{TP}{TP+FP+FN},$$

being TP, TN, FP, FN true positives, true negatives, false positives and false negatives, respectively. OA is traditional accuracy, typically overfavored in multiclass contexts given the high value of TN as the number of classes increments. For this

	Correct			Erroneous			
Scenario (total tracks)	(a)	(b)	Total	(c)	(d)	(e)	Total
Hermes (161)	103	26	129	13	10	9	32
Oxford centre (180)	87	62	149	20	8	3	31
Devil's Lake (179)	49	98	147	17	10	5	32
Kingston–1 (161)	85	53	138	12	9	2	23
Kingston–2 (87)	35	33	68	7	4	8	19

Table 3.2

Number of correctly tracked (a) pedestrians and (b) vehicles in each scenario, and amount of observation errors due to: (c) agent misclassification, (d) lost or missed tracks, and (e) false detections.

reason, SA has been increasingly used to evaluate multiclass segmentations, as in the PASCAL-VOC challenge [4]. Additionally, WSA is defined by

$$WSA = \frac{TP^*}{TP^*+FP^*+FN^*},$$

in which an assignment is now considered *positive* if the inferred label either equals the real one, or is its direct generalization; and *negative* otherwise, thus modifying the account of errors. For instance, an actual *parking* is here positively labeled as *road*, and a *pedestrian waiting zone* is correctly labeled as *sidewalk*. Note that this score does not necessarily benefit our approach, since our smoothness constraints do not award class generalization. The goal of this metric is to penalize wrong particularizations. Ground truth evaluation in Fig. 3.6a shows that WSA finds consistency in different ground truth realizations –unlike SA–, while penalizing differences more than OA.

Median filter

Median filters are the most used nonlinear filters to remove impulsive or isolated noise from an image. Their main characteristic is the preservation of sharp edges, which makes them more robust than traditional linear filters and a simple and cheap solution to achieve effective non-linear smoothing. They are commonly used for applications of denoising, image restoration, and interpolation of missing samples, all of which are applicable in our context.

We have compared the performances obtained by a median filter and by the proposed inference framework, to evaluate the contributions of taxonomy-based constraints to the smoothing task. The filter is applied for each marginal probability map $P(f_r = l), l = 1 \ldots |\mathcal{L}|$, maintaining the MRF neighborhood defined. A median-filtered labeling is performed by assigning the most probable label to each region, once the process has converged or exceeded the maximum number of iterations allowed.

[4] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Results

Fig. 3.6b shows quantitative scores for OA, SA, and WSA in the 5 scenarios, for grid sizes ranging from 40×40 to 150×150. Each plot draws the results of 4 approaches, applied to the 5 series of ground truth available. These approaches correspond to: (i) assigning labels using only observed evidence from trajectories, i.e., neglecting smoothness priors (*Initial*); (ii) using a median filter over the initial models (*Median*); (iii) applying MAP-MRF inference to the initial models (*MRF*); and (iv) applying geodesic interpolation to region models before MAP-MRF inference (*GI–MRF*).

Occasional plot oscillations are mainly due to the non-linear operation of sampling ground truth images into lattices of a concrete size. Moreover, given that the region modeling is based on point samples, augmenting the cell resolution progressively lowers the quality of the initial models, as well as the accuracy on posterior labelings. Nonetheless, it is shown that interpolation grants a performance almost invariant to the grid size used. This is emphasized in case of poor sampling, e.g, parking lots.

Table 3.3 shows numerical results for a grid of 75×75 cells, with and without considering noisy trajectories. As seen in this table, OA is excessively favored due to the high number of true negatives produced in a multiclass context, thus suggesting SA and WSA as more convenient to compare the different techniques. Particularly, WSA should be interpreted as the precaution to avoid wrong particularizations. With these metrics, experiments using geodesic interpolation and smoothness constraints practically always achieve the maximum score, whereas a median filter fails dramatically as the grid resolution augments, or in case of ill-convergence; e.g., it fails to preserve parking regions in *Kingston–2*. Additionally, it is seen that even by incorporating erroneous trajectories to the datasets, letting them be about a 20% of the total, the accuracy values remain stable.

Fig 3.7 depicts qualitative step results of the labeling process for a grid size of 75×75. For visualization purposes, results are shown within a ROI. The depicted results represent the activity of the tracked objects, rather than the visual appearance of the scenario. Instead, appearance is commonly used to guide manual labelings. We also identify an edge-effect of *Void* regions, given that control points near the edges often lack of precedent or consecutive samples to update their regions. This happens especially for vehicles, due to their higher speed and poorer sampling. Finally, cases of intense projectivity –e.g., *Devil's Lake*–, make it more difficult for the models to emphasize the presence of connected regions, thus provoking generalized smoothing.

3.6 Discussion

We have shown an effective motion-based method for automatic semantic segmentation and labeling in urban scenarios. Our approach enhances state-of-the-art on background labeling by using prior taxonomical knowledge to guide consistent inferences during labeling. In addition, it is invariant to viewpoint and of reduced computational cost, for it does not require to compute costly image descriptors.

Initial region models are learned from trajectory features, and updated as new trajectories are available. Smoothness is taken into account using a MAP-MRF inference, whose parameters are conditioned by prior taxonomical knowledge on the

		Overall accuracy (OA)				Segmentation accuracy (SA)				Weighted segmentation accuracy (WSA)			
		Initial	Median	MRF	GI–MRF	Initial	Median	MRF	GI–MRF	Initial	Median	MRF	GI–MRF
Only correct	Hermes	0.98	0.96	0.97	0.98	0.40	0.40	0.45	**0.64**	0.50	0.44	0.51	**0.77**
	Oxford Centre	0.98	0.97	0.98	0.98	0.46	0.52	0.58	**0.61**	0.65	0.66	0.75	**0.93**
	Devil's Lake	0.98	0.99	0.99	0.99	0.37	0.39	0.39	**0.44**	0.49	0.46	0.52	**0.78**
	Kingston–1	0.98	0.97	0.98	0.99	0.43	0.37	0.50	**0.66**	0.46	0.44	0.59	**0.76**
	Kingston–2	1.00	0.84	1.00	0.98	0.27	0.24	0.28	**0.56**	0.36	0.24	0.35	**0.69**
	Average	0.98	0.94	0.98	0.98	0.39	0.38	0.44	**0.58**	0.49	0.45	0.54	**0.78**
Correct and erroneous	Hermes	0.98	0.97	0.97	0.98	0.40	0.40	0.45	**0.53**	0.51	0.45	0.52	**0.78**
	Oxford Centre	0.98	0.97	0.98	0.98	0.46	0.53	0.56	**0.57**	0.66	0.68	0.76	**0.94**
	Devil's Lake	0.98	0.99	0.99	0.99	0.37	0.39	0.40	**0.43**	0.50	0.47	0.53	**0.78**
	Kingston–1	0.97	0.98	0.98	0.98	0.43	0.40	0.50	**0.65**	0.46	0.50	0.60	**0.76**
	Kingston–2	1.00	0.84	0.99	0.98	0.28	0.24	0.34	**0.55**	0.38	0.26	0.40	**0.76**
	Average	0.98	0.95	0.98	0.98	0.39	0.39	0.46	**0.55**	0.50	0.47	0.56	**0.80**

Table 3.3

Quantitative OA, SA, and WSA scores for a grid size of 75×75, without and with the presence of erroneous trajectories.

domain. The framework is scenario-independent: it has been applied to 5 datasets showing different conditions of projectivity, region content and configuration, and agent activity. We have shown step results at every stage of the process, to capture the particular contributions of each proposed task. The method has been compared to a median filter, showing its better performance on the 3 scores tested.

Further steps include extending the system to indoor scenarios. Such environments incorporate more complex semantics on agent behaviors, and present challenging tracking difficulties like occlusions or clutter, which could be solved as well with the use of domain knowledge.

Resum

En aquest capítol s'ha descrit un mètode basat en l'anàlisi de trajectòries que permet realitzar tasques de segmentació i etiquetatge de les regions semàntiques de l'escenari, de forma automàtica, per a escenaris de videovigilància de tipus urbà. La nostra proposta millora l'estat de l'art actual en etiquetatge de fons d'escena basat en moviment, pel fet que fonamenta les anàlisis en coneixement taxonòmic a priori, que guia el procés d'etiquetatge per tal de realitzar inferències consistents. Addicionalment, el mètode és invariant en quant al punt de vista i té un cost computacional reduit, per la qual cosa no requereix l'ús de descriptors d'imatge computacionalment costosos.

El mètode descrit es duu a terme fonamentalment en dues parts: primerament, les regions inicials del model semàntic s'aprenen a partir de característiques de cada trajectòria observada, i es van actualitzant automàticament a mida que es reconeix una nova trajectòria. En segon lloc, el model inicial, força sorollós, es suavitza fent servir tècniques d'inferència MAP-MRF, els paràmetres de la qual estan condicionats per coneixement taxonòmic a priori del domini en qüestió.

La proposta és independent de l'escenari. S'ha aplicat a 5 bases de dades, cadascuna d'elles amb diferents característiques de projectivitat, tipus de regions contingudes, configuració estructural d'aquestes regions, i activitats d'agents observades per les càmeres. Les dades de tots els conjunts són reals, la meitat de càmeres web públiques.

S'ha demostrat el bon funcionament del sistema per als conjunts de dades proporcionats. S'han recollit els resultats parcials per a cada etapa del procés, per tal d'entendre les contribucions particulars de cada tasca proposada. Hem comparat el nostre mètode amb tècniques tradicionals com ara el filtre de mediana, i demostrat el bon funcionament de la nostra proposta en els tres sistemes d'avaluació provats.

Les següents passes que hem considerat inclouen principalment el pas de l'avaluació de la tècnica de dominis diferents exteriors urbans a altres de diferents, com ara escenes interiors. Aquests nou domini incorpora tipus de semàntica més complexes pel que fa al comportament dels agents respecte el seu entorn, i presenta serioses dificultats quant al seguiment automàtic dels agents, com ara oclusions i *clutter*, que podrien solucionar-se fent un bon ús del coneixement de domini a priori.

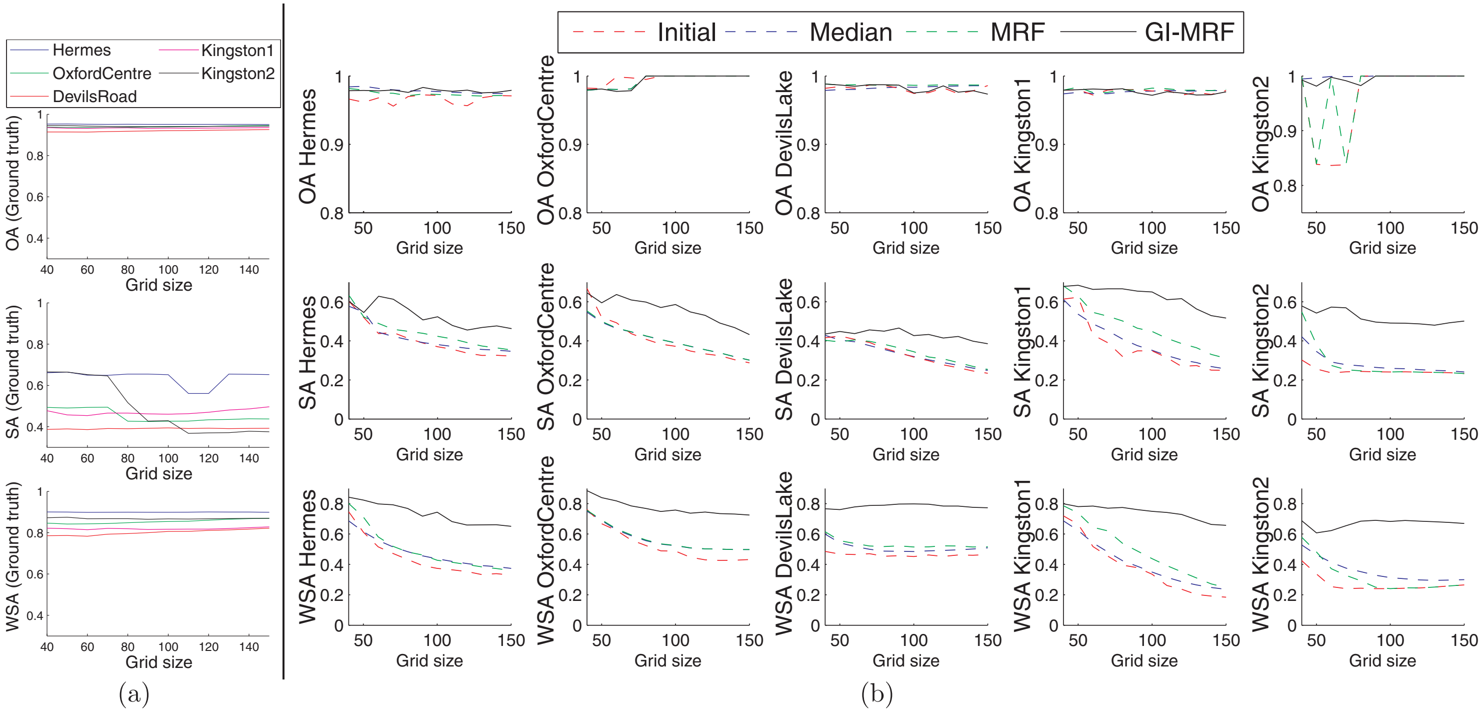


Figure 3.6: (a) Evaluation of the inter-observer variability in ground truth segmentations. (b) Statistical scores for the 5 considered scenarios. In both cases, grid sizes range from 40 × 40 to 150 × 150. More details in the text.

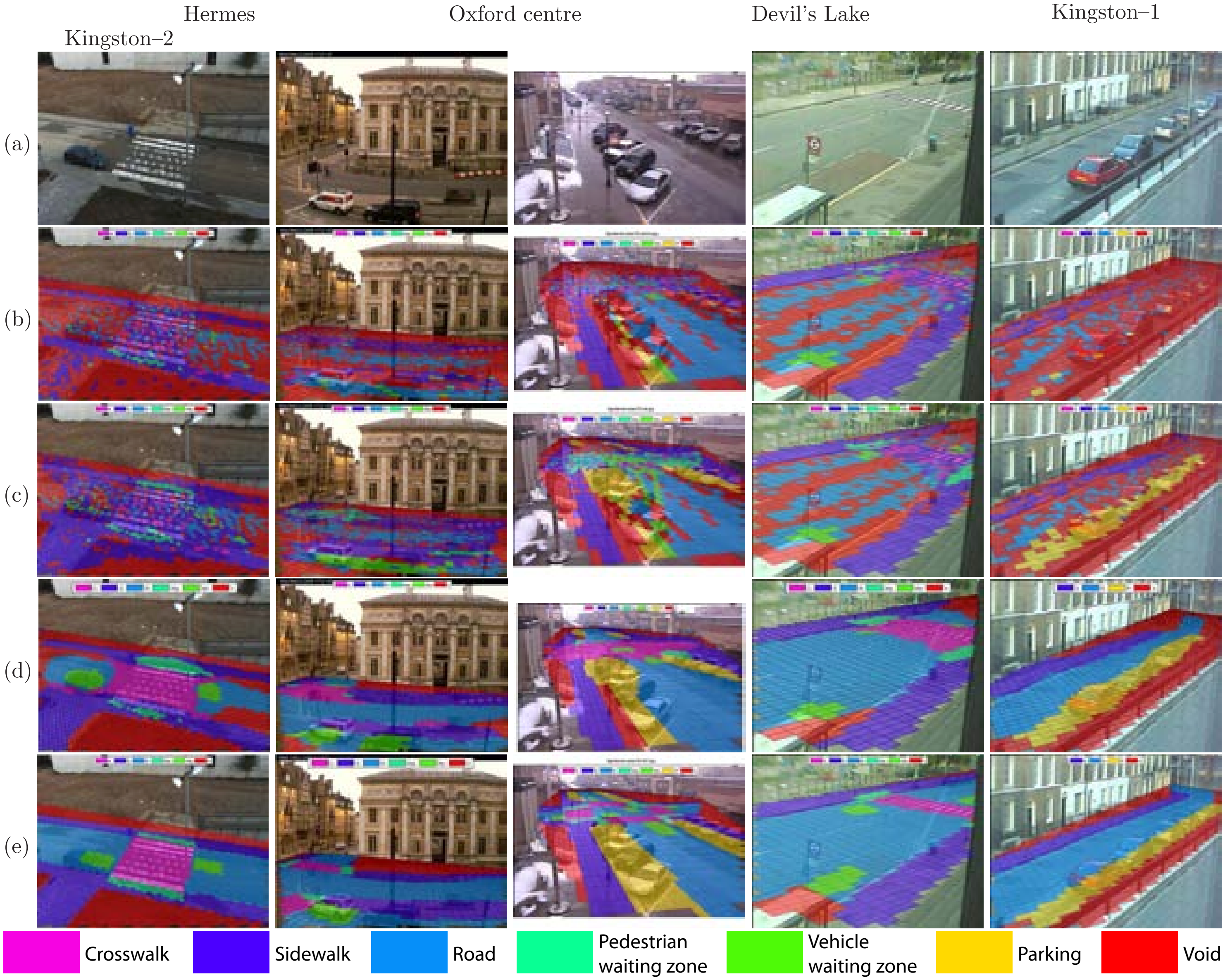


Figure 3.7: Step results of 5 region labelings for a grid size of 75 × 75: (a) original image, (b) initial labeling only based on observations, (c) initial labeling with geodesic interpolation, (d) inference labeling using both interpolation and smoothness constraints, (e) ground truth example. Best viewed in color.

Chapter 4

Ontologies for behavior modeling and interpretation

> *"Once you've seen the signs about the barn, it becomes impossible to see the barn. (. . .) We're not here to capture an image, we're here to maintain one."*
>
> *White Noise* (1985), by Don DeLillo

The aim of this stage is twofold. First, it should reason about collected visual evidence, and provide a holistic interpretation of the facts according to prior models; we will discuss how to build and apply these models. Secondly, it should articulate the semantic knowledge for rapid exploitation at any level of the system, i.e., from cameras to end-users; thus, we will also argue about the centralizing role of this semantic stage. The gaps involved in this chapter are mainly the semantic and model gap.

The field of video understanding has received much interest in recent years. In general, it aims to translate video sequences into high-level semantic concepts. This field typically requires a step of *event modeling*, which becomes central to understand video content in many applications like smart surveillance, advanced user interfacing, or semantic video indexing. However, the interpretation of visual evidence for video understanding is not trivial: as described in the introduction of this thesis, we find an inherent ambiguity between a sequence of images and its possible interpretations, the *semantic gap* [116] described in the introduction.

In order to bridge the semantic gap, it has been proved useful to rely on semantic models, which aim to detail the essential lower-level attributes of the high-level terms of interest, and restrict their applications. Among all semantic models, ontologies become especially useful, for they provide explicit structure –hierarchy, dependencies– to a set of chosen concepts, integrate them into a single repository, and enable the derivation of implicit knowledge through automated inference. Event modeling finally has to provide the formal description of ontologies and other types of semantic models, thus enabling further recognition of spatiotemporal events.

As pointed out in [74], the main general question in event modeling is *"How can the events of interest be represented and recognized?"*. Nevertheless, another important question arises prior to this one when facing a domain of interest, which sometimes is not given enough attention: *"Which semantic concepts should be chosen, in order to build the different interpretative models?"* As an answer to this question, techniques for *concept selection* are applied to facilitate the first step on the building of semantic models.

4.1 Top-down modeling for event recognition

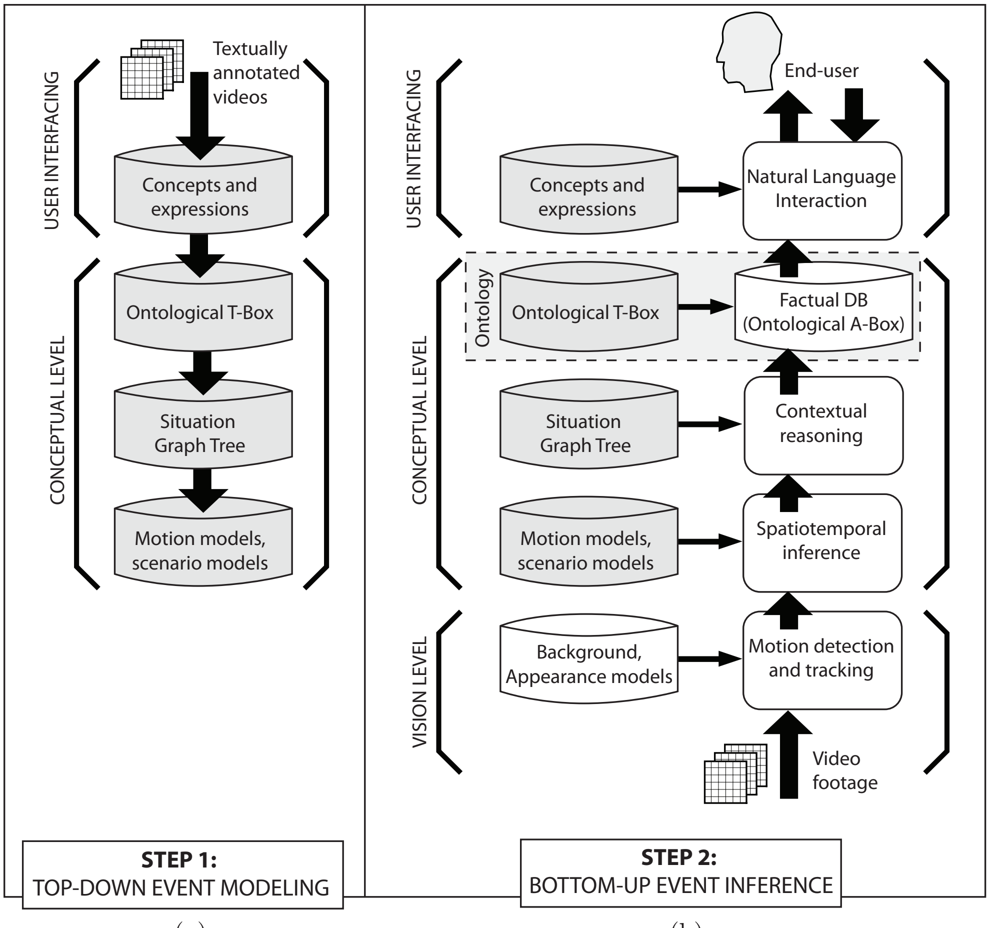
Whereas the selection of semantic concepts in event modeling is often implicit and unstructured, we suggest a guided approach based on the usage of terms within NL discourses. Our top-down method consists of having experts gather NL textual evidence from human participants, and use the implicated semantics to define ontological resources for multiple applications based on video content interpretation. The advantage of this approach is to adjust completely to the aim of the application, considering a minimal set of relevant concepts that are statistically consistent with the usual descriptions. However, we have the drawback of dealing with linguistic definitions, usually vague or imprecise, that could be difficult to match with the inference capabilities of the system.

The general architecture of the proposal is presented in Fig. 4.1. We divide the system in 3 distinguished levels devoted to visual, conceptual, and user interfacing tasks, and the presented process is as well divided in 2 steps: an initial top-down modeling of the knowledge bases guided by an expert, and a subsequent automatic, bottom-up inference by the system using the resulting event models.

The *top-down modeling* process depicted in Fig. 4.1 works as follows: first, based on several training videos, we gather event descriptions reported by a large number of non-expert users and assess the variability of these reports. The descriptions are then used to build the semantic models in a strict top-down fashion, unlike the majority of current approaches for video indexing and understanding. Top-down approaches enable an a priori selection of relevant features, which is an advantage with respect to the generic models used in bottom-up approaches, given that we define procedures that are goal-directed [87]. Our integrative architecture incorporates a large component of domain-knowledge that is managed by dedicated modules, a common characteristic of expert systems.

As a result of the top-down modeling process, a series of semantic models and knowledge bases are obtained at different stages of the system. The next step, *bottom-up inference*, automatically produces high-level interpretations of occurrences for generic image sequences in the domain. Eventually, it also facilitates different forms of user-interaction: natural language texts, query-based retrieval of information, and generation of virtual sequences.

Next sections detail how to accomplish the top-down modeling of events. The first part of this chapter describes the top-down modeling employed to address the task of knowledge management. The resulting models are later used for inferential reasoning and video understanding. The different steps include:



(a) (b)

Figure 4.1: General overview. (a) First, knowledge bases are built top-down, based on end-user event descriptions. (b) Once domain knowledge is modeled, any video in the domain can be automatically indexed for retrieval, in a bottom-up fashion.

1. building a domain ontology from NL questionnaires of event description run on several subjects,

2. contextualizing targeted events with concrete models that decompose them into simple facts, and

3. link these facts to spatiotemporal data available from tracking.

The target events to be detected in surveilled footage are typically closed and şdetermined by the purposed application. Nevertheless, assessing interpretations often becomes uncertain when dealing with complex events, leading to engineered solutions that may differ from end-user's perceptions. In order to deal with this, we have

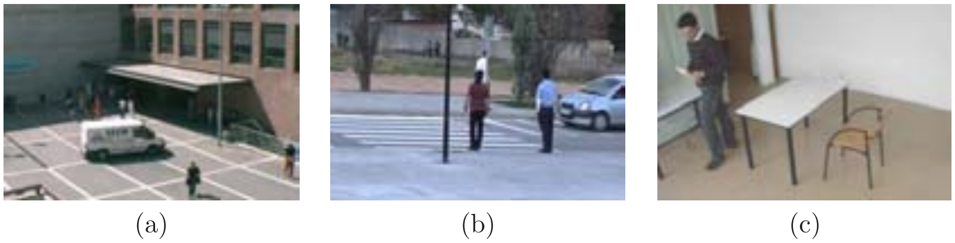


Figure 4.2: Snapshots of outdoor (a)(b) and indoor (c) video surveilled scenarios used for the ground-truth annotation of semantic evidence.

run questionnaires to identify which events are relevant to end-users in our restricted domain, in order to model them in a top-down fashion.

The ground-truth annotation of events has been extracted this way from psychophysical experiments of manual video annotation. Three scenes from indoor and outdoor scenarios have been recorded, showing different kind of interactions among people, objects, and vehicles, see Fig. 4.2. They show some complex events like stealing objects, crossing roads, waiting to cross, or getting almost run over by cars. A population of 60 English speakers were requested to visualize the videos[1]. 40 of the subjects were told to annotate at least 20 notable occurrences happening in each training sequence, the other 20 did the same for the two test sequences used for experimental results. Similar annotations were manually gathered together by experts, e.g. 'talk' – 'have a conversation' – 'discuss' → 'talk to someone'. Table 4.1 gives the frequency of common annotations for outdoor and indoor training videos. For events occurring more than once in the same video, the maximum frequency was considered.

4.2 Ontological modeling

The main motivation for the use of ontologies is to *capture the knowledge involved in a certain domain of interest*, by specifying some conventions about the content implied by this domain. Ontologies are especially used in environments requiring to share, reuse, or interchange specific knowledge among entities involved in different levels of manipulation of the information.

There exist many approaches for the ontological categorization of visually perceived events. An extensive review is done in [81], from which we remark Case Grammar, Lexical Conceptual Structures, Thematic Proto-Roles, WordNet, Aspectual Classes, and Verb Classes, which focus on the use of eventive verbs as main representative elements for classifying types of occurrences. As an extension, our approach relates each situation from an ontology with a set of required entities, which are classified depending on the thematic role they develop. The main advantage of this approach is an independency of the particularities of verbs from a concrete natural language, thus facilitating addition of multiple languages.

[1] The subjects were recruited from 5 different countries and from different age intervals: 18–25 (12%), 25–35 (66%), and over 35 (22%). They also came from different backgrounds: technical studies (27%), sciences (40%), humanities (30%), or none of the previous (3%).

Use	*Annotations for Outdoor Scenarios*
100%	leave object
100%	wait/try to cross
90%	walk in a location
86%	cross the road
84%	run off/away
84%	yield someone
80%	chase after someone
70%	pick up an object
63%	join someone at a location
60%	appear in a location
50%	steal object from someone
47%	do not allow someone to cross
44%	danger of runover

(a)

Use	*Annotations for Indoor Scenarios*
100%	pick up / retrieve object
96%	leave a location
96%	use vending machine
96%	sit down at a table
92%	talk to someone
90%	appear in a location
88%	leave object on the floor
85%	stand up
81%	shake hands with someone
69%	kick/hit vending machine
62%	carry an object
58%	go/walk to a location
50%	abandon/forget an object

(b)

Table 4.1

Most common annotations for (a) outdoor and (b) indoor scenarios, sorted by the percentage of people using them while describing the events.

The design of the ontology for the described cognitive vision system has been done putting especial effort on the definition of the knowledge base. DL allows us to structure the domain of interest by means of *concepts*, designing sets of objects, and *roles*, denoting binary relations between concept instances [8]. Specifically, our domain of interest is represented by a knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, which contains two different types of knowledge:

- A TBox $\mathcal{T}$ storing intensional knowledge, i.e. a set of concept definitions which classify the terminological information of the considered domain. In practice, we split the terminology into several TBoxes (i.e. taxonomies), according to the semantic nature of the participants for each set. Some of the main important sets are Event-TBox (see Table 4.2), Entity-TBox, and Descriptor-TBox (see Table 4.3).

- An ABox $\mathcal{A}$ storing assertional knowledge, i.e. factual information concerning the world state and the set of individuals which can be found in it. This extensional knowledge will be first instantiated by reasoning and inference stages dealing with First-Order Logic, and then introduced into the relational database by means of concept assertions, e.g. `pedestrian(Agent1)` and role assertions, e.g. `enter(Agent2, Crosswalk)`

The ontology language we use has been restricted to the $\mathcal{SHIF}$ family (a.k.a. DL-Lite), which offers concept satisfiability and ABox consistency to be log-space computable, thus allowing the relational database to handle in practice large amounts of data [1].

An ontology of events has been created out of the results provided. Each annotation incorporates, explicitly or implicitly, the semantic context required to model an event, by means of a series of concepts that have been structured in 3 categories: events, entities, and constraints. The *Event* concepts identify the occurrence described, and are organized from simple to complex as (i) spatiotemporal inferences from tracking, (ii) interactions among entities, and (iii) interpretations of complex events in specific contexts. *Entity* concepts determine the nature of the participants in the event, which can be agents, objects, or locations. Finally, *Constraint* concepts account for the roles that entities are required to satisfy within an event, i.e., the list of agents, patients, locations, or objects needed. All these concepts are classified in taxonomies and together conform the terminological part of the ontology, the so-called TBox $\mathcal{T}$ [47]. Table 4.5 reports how the annotated events are used to build the TBox of the ontology: the entities required by each event are identified, and related to the particular event by means of constraints, which give additional information on the type of relationship held with each of the entities.

Apart from $\mathcal{T}$, the ontology also incorporates an ABox $\mathcal{A}$ storing concept instances, i.e., factual information regarding the world state and the individuals existing on it [47]. Once the abstract events, constraints, and entities are satisfied for a certain world state, these concepts are instantiated into the factual database as *Facts*, *Constraint instances*, and *Entity instances*, respectively. For example, for the *theft* event in Table 4.7, the ontology requires a thief, *isAgent(Pedestrian)*, a victim, *has_agent_interaction(Pedestrian)*, and a stolen item, *has_object_interaction(Object)*, in this case fulfilled by instances *ped2*, *ped1*, and *obj1*, respectively.

In the end, the domain of interest is formally represented by a knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, the factual database, which includes both the concepts and their instances. Fig. 4.11 gives a concise view of the factual database implemented: the abstract concepts are *Events*, *Entities*, and *Constraints* that state which entities are needed for which events. On the other hand, instances for these 3 types of concepts are stored in the 3 other tables: *Entity instances* list appearing entities, *Facts* are detected occurrences of events, and *Constraint instances* link ones to the others.

Talmy organizes conceptual material in a cognitive manner by analyzing what he considers most crucial parameters in conception: space and time, motion and location, causation and force interaction, and attention and viewpoint [122]. For him, semantic understanding involves the combination of these domains into an integrated whole. Our classification of situations (i.e. the Event-TBox, the central element in our ontology) agrees with these structuring domains: We organize semantics in a linear fashion, ranging from structural knowledge in vision processes (quantitative pose vectors) to uncertain, intentional knowledge based on attentional factors (high-level interpretations). It is structured as follows, see Table 4.4:

- At the lowest level we consider *spatiotemporal data* retrieved from motion tracking. Here we include positions, orientations, or static configurations –poses, facial meshes– at given time-steps. No class is created for them, since semantics is only present in form of structural information by means of quantitative values.

- The *Status* class contains metric-temporal knowledge, based on the information provided by the considered trackers: body, agent, and face. Its elements rep-

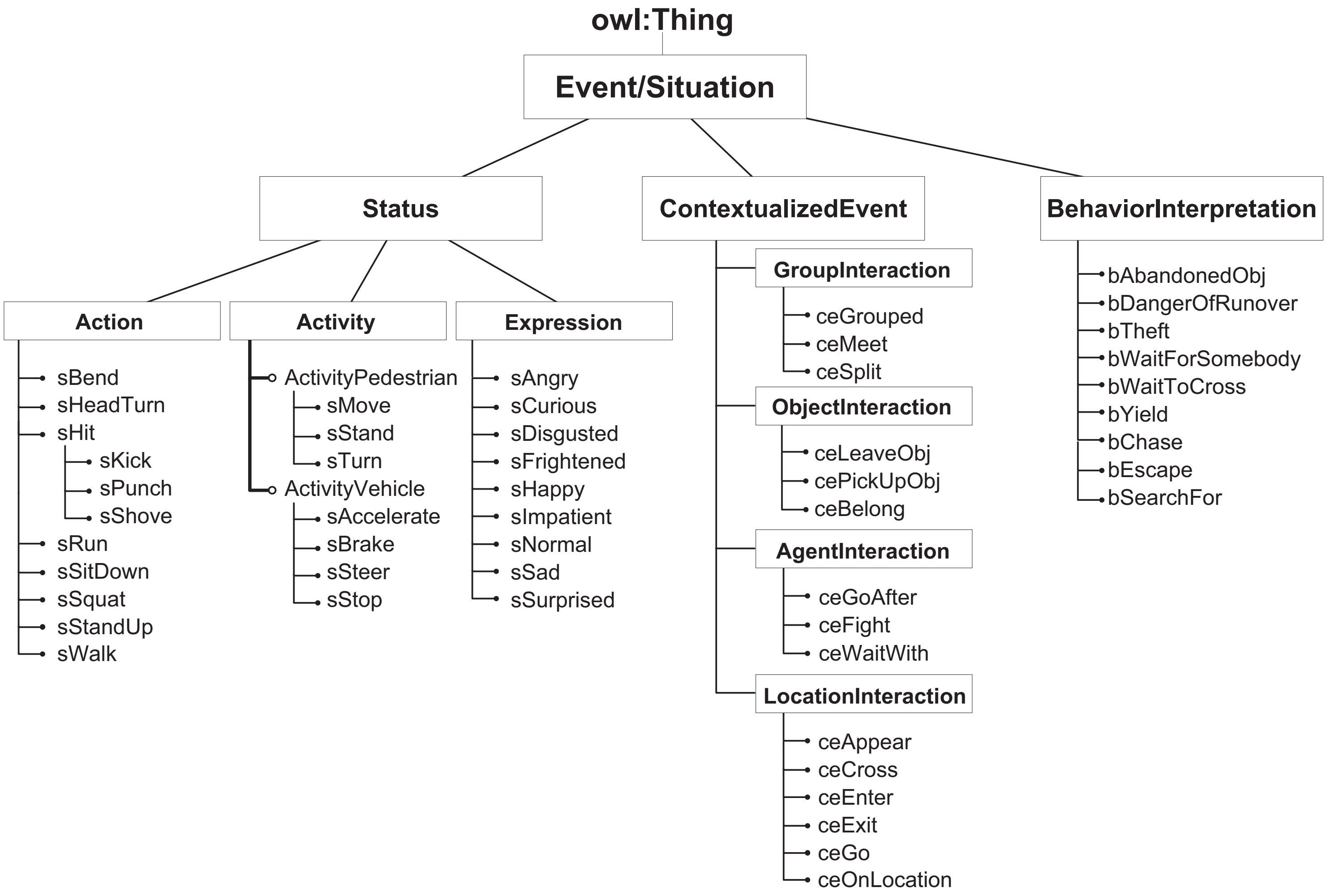


Table 4.2
Taxonomy containing some concepts from the Event-TBox.

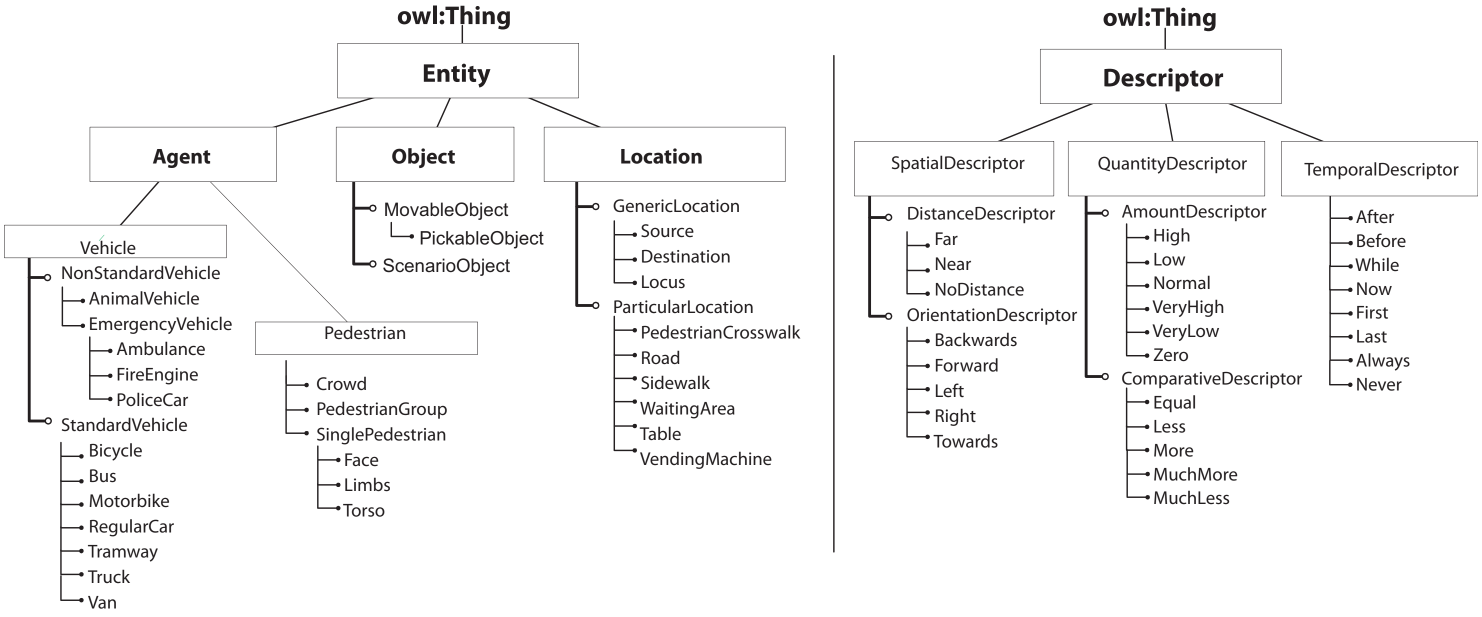


Table 4.3

Taxonomies showing highlighted concepts from the Entity-Tbox (left) and the Descriptor-TBox (right).

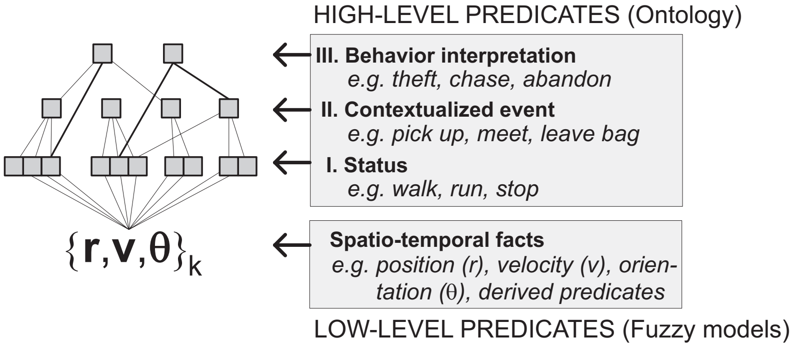

HIGH-LEVEL PREDICATES (Ontology)
III. Behavior interpretation
e.g. theft, chase, abandon
II. Contextualized event
e.g. pick up, meet, leave bag
I. Status
e.g. walk, run, stop
$\{r, v, \theta\}_k$
Spatio-temporal facts
e.g. position (r), velocity (v), orientation (θ), derived predicates
LOW-LEVEL PREDICATES (Fuzzy models)

Table 4.4

A knowledge-based classification of human behaviors in urban contexts. High-level events are conjunctions and sequences of lower-level events. This terminology structures the Event-TBox and guides interpretation.

resent dynamic interpretations of the spatial configurations and trajectories of the agents. Some examples include to detect that a pedestrian is turning left, or that a car is accelerating.

- The *ContextualizedEvent* class involves semantics at a higher level, now considering interactions among semantic entities. This knowledge emerges after contextualizing different sources of information, e.g. 'sit down'–'bus stop', or 'wave hand'–'open mouth', that allows for anticipation of events and reasoning of causation.

- Finally, the *BehaviorInterpretation* class specifies event interpretations with the greatest level of uncertainty and the larger number of assumptions. Intentional and attentional factors are considered, here the detection of remarkable behaviors in urban outdoor scenarios for surveillance purposes.

This classification of knowledge will guide the process of interpretation. It can be seen that this proposal takes into account all levels of extraction of visual information which have been thought for the cognitive vision system –i.e. agent, body, face, and relation with other detected objects, agents, and events–, and also suggests a proper way of managing the different stages of knowledge. This categorization considers the relevance of the retrieved information, some hierarchical degrees of perspective, and also the level of subjectiveness required for a scene interpretation, as will be explained in the following sections.

As stated in [99], changes in the topology and distribution of the ontological knowledge do not hold special significance. What is much more crucial is to focus on coverage, i.e., to find a suitable grain size of semantic representations to fulfil a concrete application. The main idea is to model high-level, more subjective behaviors in a way such as they are not wrongly extended to general situations, while not

	User annotation	Event	Entities	Constraints
Outdoor	*wait to cross*	bWaitToCross	Pedestrian	is_agent
			Location	hasLocationInteractionWith
	danger of runover	bDangerOfRunover	Vehicle	is_agent
			Pedestrian	hasPatientInteractionWith
Indoor and outdoor	*leave a location*	ceExit	Agent	is_agent
			Location	hasLocationInteractionWith
	pick up object	cePickUpObj	Pedestrian	is_agent
			PickableObject	hasObjectInteractionWith
	meet with someone	ceMeet	Pedestrian	is_agent
			Pedestrian	hasPatientInteractionWith
			Location	hasLocationInteractionWith
	abandon/forget object	bAbandonedObj	PickableObject	isObject
			Location	hasLocationInteractionWith
	steal object from someone	bTheft	Pedestrian	is_agent
			PickableObject	hasObjectInteractionWith
			Pedestrian	hasPatientInteractionWith

Table 4.5

List of examples on how user annotations are used to interrelate concepts from the TBox $\mathcal{T}$.

requiring excessively detailed information for deductions. That is why the described approach has been designed to work at different levels of representation regarding the generality of situations, and the reason for the general architecture to have been conceived in terms of collaborative modules.

4.3 Contextual modeling

At this point, the ontology already states which elements are required by each event, but we still need to model the domain-specific context in which an event occurs. As stated before, events are *situated* in their context by means of SGTs.

An independent stage is implemented to achieve effective modeling of behaviors and complex situations. The concurrence of hundreds of conceptual predicates makes necessary to think of a separate module to deal with new semantic properties at a higher level: some guidelines are needed to establish relations of cause, effect, precedence, grouping, interaction, and in general any reasoning performed with time-constrained information at multiple levels of analysis. Thus, this part of the modeling deals with the contextualization and interpretation of events.

Conceptual predicates are widely used in model-based approaches in order to instantiate and infer pieces of knowledge, in systematic procedures [95, 76, 17]. In our case, conceptual predicates enable flexible reasoning from the motion data, and the inclusion of this information into the ontology. On the other hand, if we had to incorporate all the pieces of information needed to recognize events as those shown, it would result in a combinatorial explosion of instances in the ontology. For example, it could be instantiated that *a person is far from a table*, *far from a door*, *close to a machine*, *moving slow*... taking into account the assertions of all entities and all possible relationships. To minimize this problem, we distinguish between two different types of predicates: low-level and high-level.

We use low-level predicates to state the most basic spatiotemporal properties, directly defined by fuzzy motion models; for example, the distance between two tracked objects is described as `far`, `medium`, or `close` using the predicate `has_distance(Entity, Descriptor)`. Similarly, low-level predicates like `has_speed` or `similar_direction` are modeled as well. A fuzzy metric-temporal reasoner is specifically used to reason about these low-level facts, and extract higher-level information.

On the other hand, we define a high-level predicate for each event included in the Event-TBox, e.g.,

```
bAbandonedObj  →  bAbandonedObj (PickableObject, Location)
    bTheft     →  bTheft (Pedestrian, PickableObject, Pedestrian)
```

Each one of these predicates maintains semantic relationships among a set of entities –and possibly, descriptors–, and these relations are explicitly expressed and stored in the ontology. Since the amount of high-level predicates is much less than the number of low-level ones, the computational load is efficiently shared. Then, though, another question arises: *"How can we express semantic concepts in terms of tracking output?"*

The tool chosen to articulate high-level predicates in terms of low-level ones is the SGT, see [6, 43]. An SGT is a hierarchical classification tool used to describe behavior of agents in terms of situations they can be in. These trees contain a-priori knowledge about the admissible sequences of occurrences in a defined domain. Basing on deterministic models built upon elements of the ontology, they explicitly represent and combine the specialization, temporal, and semantic relationships of the conceptual facts which have been asserted.

The semantic knowledge related to any agent at a given point of time is contained in a *situation scheme*, which constitutes the basic component of a SGT, see Fig. 4.3. A situation scheme can be seen as a semantic function that evaluates an input consisting of the conjunction of a set of conditions –the so-called *state predicates*–, and generates logic outputs at a higher level –the *action predicates*– once all the conditions are asserted. Here, the action predicate is a `note` method which generates a semantic annotation in a language-oriented form, containing fields related to thematic roles such as *Agent*, *Object* or *Location*, which refer to participants of the Entities-TBox in the ontology.

On the other hand, the temporal dimension of the situation analysis problem is also tackled by the SGT. As seen in Fig. 4.4, the situation schemes are distributed along the tree-like structure by means of three possible directional connections, the *particularization*, *prediction*, and *self-prediction edges*. Particularization edges allow to instantiate more specific situations once the conditions of a general situation have been accomplished. On the other hand, prediction edges inform about the following admissible states within a situation graph from a given state, including the maintenance of the current state by means of self-prediction edges. Thus, the conjunction of these edges allow defining a map of admissible paths through the set of considered

ID	High-level predicate	Temporal decomposition	
①	ceLeaveObj(Object, Agent)	t_0 :	ceSplit(Agent,Object) ∧ has_speed(Object, zero)
②	bAbandonedObj(Object,Agent)	t_0 :	ceLeaveObj(Object, Agent)
		t_1 :	has_distance(Agent, Object, far) ∧ has_speed(Object, zero)
③	cePickUpObj (Agent, Object)	t_0 :	bAbandonedObj(Object, Agent)
		t_1 :	ceGrouped(Agent, Object) ∧ has_speed(Object, V) ∧ is_not(V, zero)
④	sStand (Pedestrian)	t_0 :	has_speed (Pedestrian,V) ∧ is_not (V, zero)
		t_1 :	has_speed (Pedestrian, zero)
⑤	sRun (Pedestrian)	t_0 :	has_speed(Pedestrian, high)
⑥	bTheft(Agent,Object,Agent)	t_0 :	ceSplit(Pedestrian,Object)
		t_1 :	object_alone(Object)
		t_2 :	agent_near_obj(Pedestrian1,Pedestrian2) ∧ Pedestrian1 <> Pedestrian2

Table 4.6

To model SGTs, high-level events are decomposed into conjunctions of simpler events that are temporally chained. Obtained decompositions are then merged into a single tree of situations for each agent type.

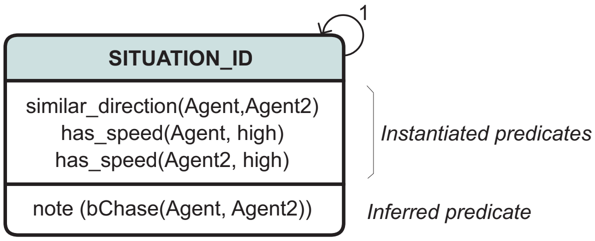


Figure 4.3: Situation scheme from a SGT. When a set of low-level predicates –the conditions– are instantiated, a high-level predicate is generated.

situations. A part of a basic SGT is shown in Fig. 4.10, which illustrates a model to identify situations such as an abandoned object or a theft.

As previously shown in Fig. 4.7, the behavioral model encoded into a SGT is traversed and converted into logical predicates, for automatic exploitation of its situation schemes. Once the asserted spatiotemporal results are logically classified by the SGT, the most specialized application-oriented predicates are generated as a result. These resulting high-level predicates are indexed with the temporal interval in which they have been the persistent output of the situational analysis stage. As a result,

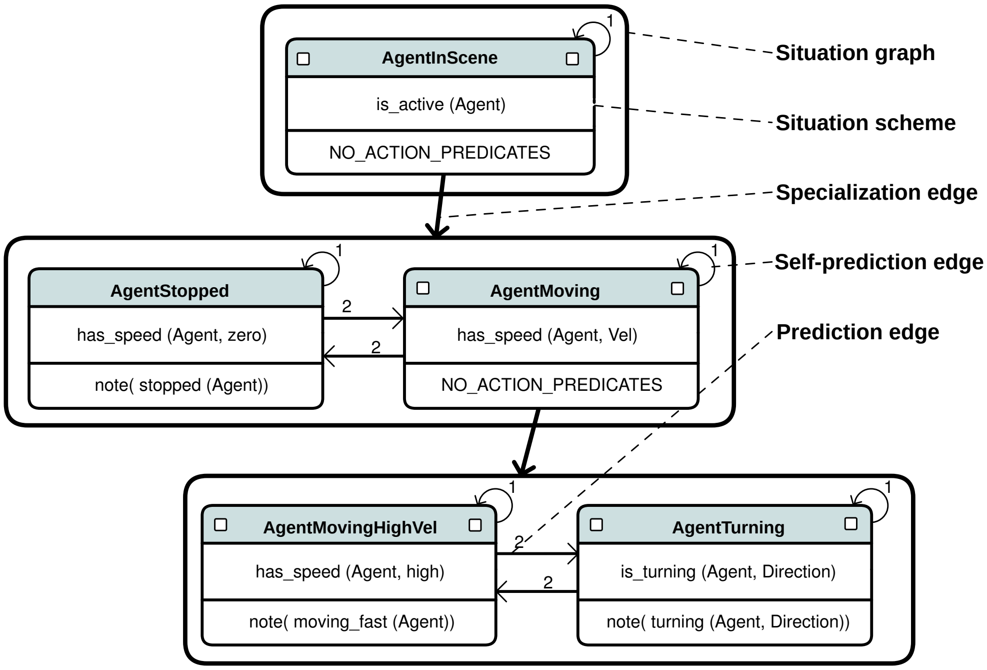


Figure 4.4: Naive example of a SGT, depicting its components. Specialization edges particularize a general situation scheme with one of the situations within its child situation graph, if more information is available. Prediction edges indicate the situations available from the current state for the following time-step; in particular, self-prediction edges hold a persistent state.

the whole sequence is split in time-intervals defined by these semantic tags. These intervals are individually cohesive regarding their content.

By describing situations as a conjunction of low-level conditions, and interrelating those situations among them using prediction and specialization edges, the *contextualization* stage described in the taxonomy of situations is accomplished. On the other hand, since the high-level action predicates are modeled depending on the application, a particular attentional factor is established over the universe of occurrences, which can be understood as the *interpretation* of a line of behaviors, for a concrete domain and towards a specific goal.

The results obtained from the behavioral level, i.e. the annotations generated by the situational analysis of an agent, are actually outputs of a process for *content detection*. From this point of view, an SGT would contain the classified collection of all possible domain-related semantic tags to be assigned to a video sequence. In addition, the temporal segmentation of video is also achieved: since each high-level predicate is associated with the temporal interval during which it has been generated, a video sequence can be split into the time-intervals which hold a permanent semantic tag. Some experimental results regarding situational analysis are presented in Section 4.7.

An SGT defines the universe of possible situations in which an agent can partici-

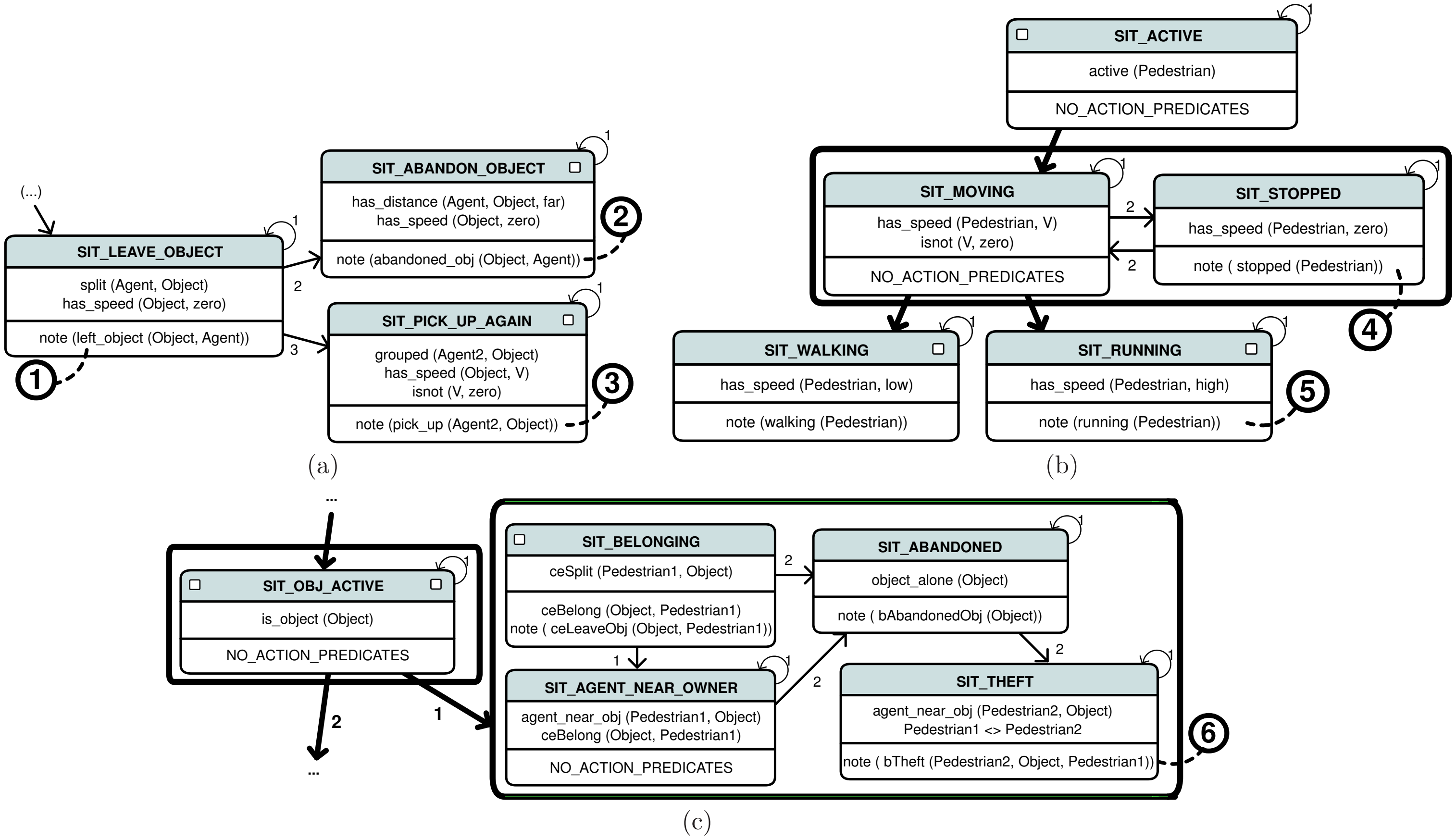

(...)
SIT_LEAVE_OBJECT
split (Agent, Object)
has_speed (Object, zero)
note (left_object (Object, Agent))
1
SIT_ABANDON_OBJECT
has_distance (Agent, Object, far)
has_speed (Object, zero)
note (abandoned_obj (Object, Agent))
2
SIT_PICK_UP_AGAIN
grouped (Agent2, Object)
has_speed (Object, V)
isnot (V, zero)
note (pick_up (Agent2, Object))
3
(a)
SIT_ACTIVE
active (Pedestrian)
NO_ACTION_PREDICATES
SIT_MOVING
has_speed (Pedestrian, V)
isnot (V, zero)
NO_ACTION_PREDICATES
SIT_STOPPED
has_speed (Pedestrian, zero)
note (stopped (Pedestrian))
4
SIT_WALKING
has_speed (Pedestrian, low)
note (walking (Pedestrian))
SIT_RUNNING
has_speed (Pedestrian, high)
note (running (Pedestrian))
5
(b)
...
SIT_OBJ_ACTIVE
is_object (Object)
NO_ACTION_PREDICATES
...
SIT_BELONGING
ceSplit (Pedestrian1, Object)
ceBelong (Object, Pedestrian1)
note (ceLeaveObj (Object, Pedestrian1))
SIT_ABANDONED
object_alone (Object)
note (bAbandonedObj (Object))
SIT_AGENT_NEAR_OWNER
agent_near_obj (Pedestrian1, Object)
ceBelong (Object, Pedestrian1)
NO_ACTION_PREDICATES
SIT_THEFT
agent_near_obj (Pedestrian2, Object)
Pedestrian1 <> Pedestrian2
note (bTheft (Pedestrian2, Object, Pedestrian1))
6
(c)

Figure 4.5: SGT mechanisms to situate events in a context: (a) temporal prediction and (b) specialization. These SGTs incorporate the decompositions shown in Table 4.6. A part of an SGT used in outdoor scenes is shown in (c).

pate. Each situation scheme evaluates a set of conditions in form of atomic predicates and reacts when all of them are asserted. In our case, reactions are `note` commands that produce the linguistic-oriented event indexes seen and facilitate NL-based retrieval [95]. Fig. 4.5(a) and (b) show parts of SGTs that exemplify their basic mechanisms to contextualize: situations are hierarchically nested from general to specific by means of *specialization* edges forming a tree, and sequentially connected by unidirectional *prediction* edges producing graphs within the tree. Self-prediction edges hold a current situation until any continuing situation applies. This scheme recurrently decomposes the evaluation of complex facts into series of low-level facts, which need to be asserted sequentially.

Carrying on the top-down modeling of semantic events, we build SGTs to define a priori the situations agents can be in. To do so, complex actions are decomposed in a combination of simpler events that are sequentially connected in time. Table 4.6 details the decomposition of the situations `left_object`, `abandoned_object`, `pick_up`, `stopped`, and `running`. It can be observed that many elements in the various decompositions are common, and thus can be merged in a single SGT. Simpler events are recursively decomposed until reaching to a combination of mere spatiotemporal descriptions. Decompositions of events like the ones shown in Table 4.6 generate the SGTs shown in Fig. 4.5(a) and (b). More complex events are also possible: for example, by combining actions like leave object, get close, pick up, and run, a *theft* event can be modeled, as shown in Fig. 4.5(c). Extra events are sometimes included into the ontology for better definition of a particular context, e.g. for the event *belongs_to*.

The role of SGTs in the overall scheme is twofold: on the one hand, they help understanding the full picture of a scene by assessing high-level interpretations from concrete pieces of information. And on the other hand, SGTs make it possible to distrust or simply neglect certain frames when the position of a target suddenly changes to a far distant location, e.g. if the tracker freezes for a while. These and similar situations make them a suitable tool to partially bridge both semantic and sensory gaps in our domain.

The current implementation of the SGT only asserts those predicates with highest confidence values, which unfits the system to handle multiple valid hypotheses at the same time, but in exchange avoids a combinatorial explosion of solutions. Only one event annotation is produced by the SGT per frame and tracked agent, which allows us to associate each predicate with an interval of validity, and build a history of events related to each detected object. When an alarm is missed at the vision level, an SGT instantiates the most specific of the events in the graph given the state conditions available. The more levels we define in the hierarchy, the more robust the system is in front of lacking information, but the computational cost increases.

4.4 Spatiotemporal modeling

The last conceptual task involves describing the multiple atomic events used in the SGTs in terms of low-level information provided by the motion trackers. To do so, a set of basic spatiotemporal rules are defined for the domain, focusing on general rather than particular contexts.

The acquisition of visual information produces an extensive amount of geometric data, considering that computer vision algorithms are applied continuously over the recordings. Such a large collection of results turns out to be increasingly difficult to handle. Thus, a process of abstraction is needed in order to extract and manage the relevant knowledge derived from the tracking processes. The question arises how these spatiotemporal developments should be represented in terms of significance, also allowing further semantic interpretations. Several requirements have to be accomplished towards this end [48]:

1. Generally, the detected scene developments are only valid for a certain time interval: the produced statements must be updated and time-delimited.

2. There is an intrinsic *uncertainty* derived from the estimation of quantities in image sequences (i.e. the sensory gap), due to the stochastic properties of the input signal, artifacts during the acquisition processes, undetected events from the scene, or false detections.

3. An abstraction step is necessary to obtain a formal representation of the visual information retrieved from the scene.

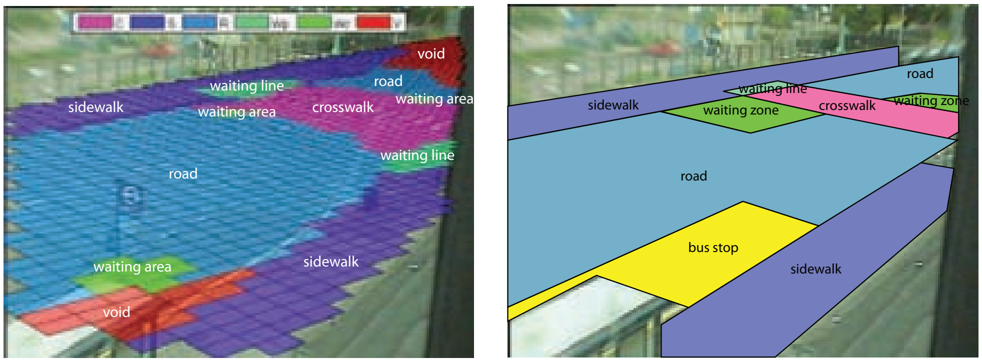
4. This representation has to allow different domains of human knowledge, e.g. analysis of human or vehicular agents, posture recognition, or expression analysis, for an eventual semantic interpretation.

FMTL has been conceived as a suitable mechanism to solve each of the aforementioned demands [112]. It is a rule-based inference engine in which conventional logic formalisms are extended by a temporal and a fuzzy component. This last one enables to cope with uncertain or partial information, by allowing variables to have degrees of truth or falsehood. The temporal component permits to represent and reason about propositions qualified in terms of time. These propositions are represented by means of *conceptual predicates*, whose validity is evaluated at each time-step.

All sources of knowledge are translated into this logic predicate formalism for the subsequent reasoning and inference stages. One of these sources is given by the motion trackers in form of agent status vectors, which are converted into `has_status` conceptual predicates [10]:

$$t \quad ! \quad has\_status(agent, x, y, \theta, v) \tag{4.1}$$

These predicates hold information for a global identification (instance id) of the agent ($agent$), his spatial location in a ground-plane representation of the scenario (x, y), and his instantaneous orientation (θ) and velocity (v). A *has_status* predicate is generated at each time-step for each detected agent. In addition, certain atomic predicates are generated for identifying the category of the agent, e.g. `pedestrian(Agent)` or `vehicle(Agent)`. The resulting categories are selected from primitives found in the Entity-TBox. Similarly, the segmented regions from the scenario are also converted into logic descriptors holding spatial characteristics, and semantic categories from the



(a) (b)

Figure 4.6: A conceptual modeling of the tackled scenario, either (a) automatically learned or (b) manually defined, is useful to derive high-level inferences.

Location-TBox are assigned to them:

$$\begin{gathered}\texttt{point (14, 5, p42)}\\ \texttt{line (p42, p43, l42)}\\ \texttt{segment (l31, l42, lseg\_31)}\\ \texttt{crosswalk\_segment (lseg\_31)}\end{gathered}\tag{4.2}$$

As detected entities are automatically classified by the motion trackers, also assigning concepts from the Location-TBox to regions of the scenario can be well accomplished in an automatic manner, as seen already in Chapter 3: each instance holds series of semantic properties, being these elements from the ABox, which can relate the instance to a particular concept after a classification process. Therefore, only methods for the obtention of semantic features are required, which can be based upon the analysis of trajectories.

The identification of semantic regions in a scenario provides *conceptual scene models* that make it possible to derive richer inferences of the observed visual data. Such models can be defined either manually or automatically, see Fig. 4.6. Automatic modeling captures the practical boundaries and limits of each semantic region, and requires none or few supervision, but may contain errors. Manual modeling, on the other hand, allows experts to describe regions with richer expressions, focused to applications of interest –e.g., linguistic descriptions–, and in a completely controllable manner, thus avoiding wrong or noisy interpretations like those sometimes produced by unsupervised procedures.

The abstraction process is thus applied over the information obtained both from the scenario and from the agents, i.e. the categorized segments from the considered location and the agent status vectors generated. Quantitative values are converted into qualitative descriptions in form of conceptual predicates, by adding fuzzy semantic parameters from the Descriptor-TBox such as *close*, *far*, *high*, *small*, *left*, or *right*. The addition of fuzzy degrees allows to deal with the uncertainty associated to visual acquisition processes, also stating the goodness of the conceptualization. Fig. 4.9 gives an example for the evaluation of a `has_speed` predicate from an asserted `has_status` fact. The conversion from quantitative to qualitative knowledge is accomplished by incorporating domain-related models to the reasoning system [95]. Hence, new inferences can be performed over an instantaneous collection of conceptual facts, enabling the derivation of logical conclusions from the assumed evidence. Higher-level inferences progressively incorporate more contextual information, i.e. relations with other detected entities in the scenario. This spatiotemporal universe of basic conceptual relations supplies the dynamic interpretations which are necessary for detecting *events* within the scene, as described in the taxonomy.

We refer to those predicates expressing uniquely spatiotemporal developments as low-level predicates. More specifically, low-level predicates facilitate a schematic representation of knowledge that is time-indexed and incorporates uncertainty. Hence, all those concepts in the Event-TBox which can be inferred only using these constraints are enclosed under this category. Low-level predicates are not only atomic: they can be generated as a result of temporal-geometric considerations. Next example shows an FMTL inference rule for the low-level predicate `similar_direction(Agent,Agent2)`:

```
always(similar_direction(Agent, Agent2):-
        has_status(Agent,_,_,_,Or1,_),
        has_status(Agent2,_,_,_,Or2,_),
        Dif1 is Or1 - Or2,
        Dif2 is Or2 - Or1,
        maximum(Dif1, Dif2, MaxDif),
        MaxDif < 30 ).
```

Hence, the FMTL reasoner engine converts geometric information into qualitative knowledge that is time-indexed and incorporates uncertainty. Note that FMTL rules are defined generally for the domain, and not dependent on particular scenes: only the semantic zones must be modeled for a new scenario. This way, the models are extensible and tracking information is easily conceptualized and forwarded to the upper levels discussed.

Fuzzy motion models are the last step of the top-down modeling process. Next section tackles the inverse approach, in which motion data from is analyzed in a bottom-up fashion, thus enabling a series of interesting applications.

4.5 Bottom-up event interpretation

Once the models have been designed top-down, the system performs *bottom-up event inference* on new image sequences. This process aims to automatically formulate interpretations of the new events observed, in form of semantic predicates. The interpretation relies on the designed models to guide the conversion from visual to semantic information.

The complete bottom-up process is represented schematically in Fig. 4.7. Video footage is firstly processed by motion trackers, which simultaneously track multiple targets in unconstrained and dynamic open-world scenarios. In our experiments, the detection of targets follows a statistical background-subtraction approach based on color and intensity cues [43]. Subsequently, the object trackers provide instantaneous target states over time, including quantitative data (e.g. velocity, size) and qualitative information (e.g. occlusions, groupings, splits, target births and deaths). Enhanced details and additional information can be found in [109]. As a result of this stage, a series of quantitative predicates are generated for each frame, such as

```
has_status(agent\_3, 2.52, 2.00, 160.44, 1.09)
has_status(agent\_2, 7.48, 6.12, 210.42, 0.78)
```

where the different values represent the (x, y) position, the degree of orientation, and the velocity that have been determined by the tracking procedures, respectively.

Secondly, a scene model is provided in order to conceptualize the spatial regions of the scenario, by putting the spatial information from motion trackers into context. This way, the spatial ground-plane coordinates (x, y) of each detected agent are assigned to regions having a priori semantic features, such as crosswalks or sidewalks for outdoor sequences, or tables or vending machines for indoor sequences, see Fig. 4.8. From this second stage, predicates of the form `in_crosswalk_segment(Agent_1)` or `in_front_of(Agent_2,vending_machine_segment)` are produced.

The third step of the process involves applying generic motion models for extended reasoning. Particularly, we are interested in the conceptualization of numerical spatiotemporal data from tracking, such as measures of velocity and orientation for the detected agents. To this end, these quantitative values are mapped to fuzzy constraints, see Fig. 4.9, so that we also preserve the uncertainty associated to the measures. New low-level predicates are generated as a result, e.g., `has_speed(Agent_1, low)` or `has_distance(Agent_2, table_2, close)`. Each of these predicates comes weighted by a degree of validity, which states the confidence on the fact according to the models. These primary facts are instantiated for each time frame in the F-Limette reasoning engine, enabling further inference of knowledge.

As described in the previous section, fuzzy models are not only used to convert from quantitative to qualitative values, but they also facilitate direct inferences coming from low-level predicates. For instance, `accelerate(Agent_1, value)` is estimated by these models using 3 consecutive values of position over time. Similarly, we can deduce whether an agent remains in the same position for a long time, or whether it follows the same direction of another agent, as also exemplified in the previous section.

In order to detect events of higher semantics, more complex patterns need to be identified. SGTs are incorporated at this point, identifying sequential patterns of asserted conditions and generating interpretations –in form of high-level predicates– as a result. The application of SGTs is done by means of a *traversal*, which evaluates the instantaneous database of FMTL facts at each frame, and tries to ascertain the conditions of the graph from a starting situation scheme. When all conditions are asserted in a situation scheme, a reaction (high-level) predicate is generated and added to the database; subsequently, a predicted, self-predicted, or specialized situation is tested in order to progress within the situation analysis. Reaction predicates are, in our case, `note` actions that state the contained predicate as an interpretation for that time step. If a condition is not accomplished, the process starts from zero.

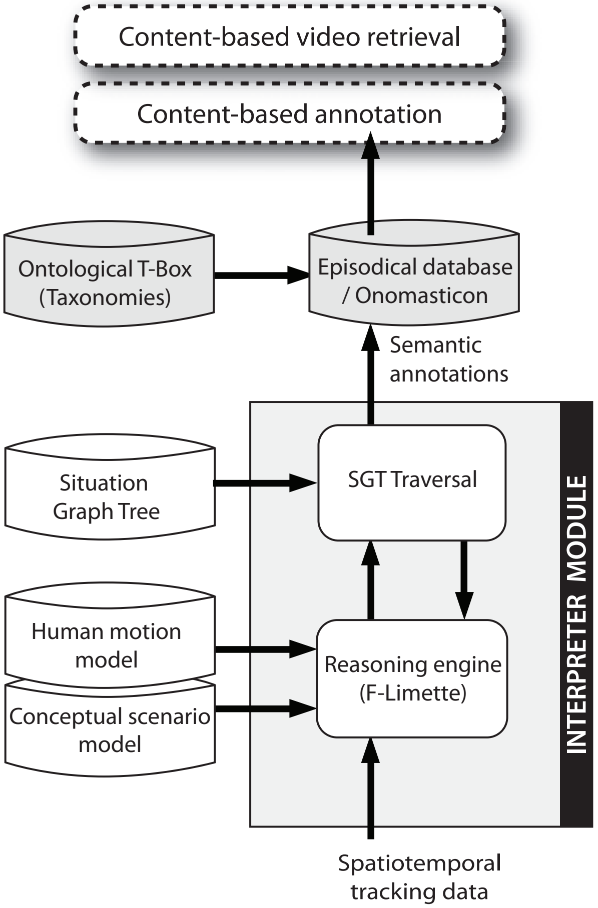
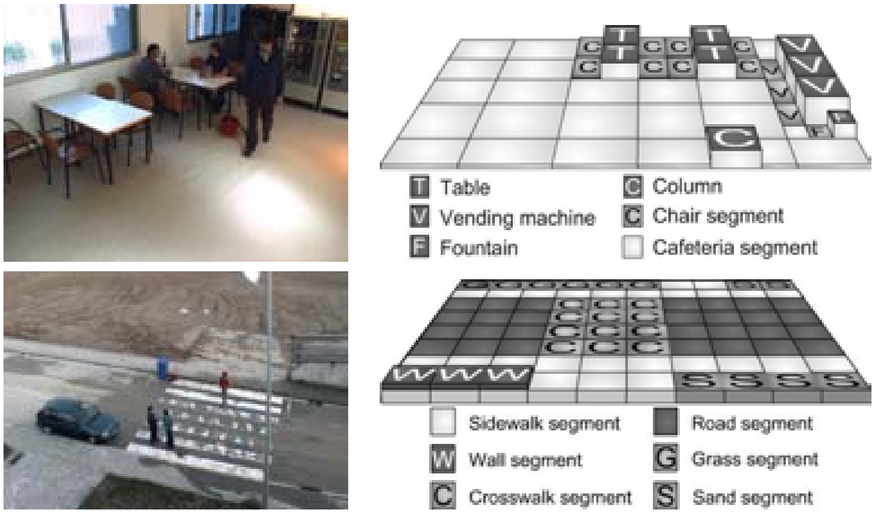

Content-based video retrieval
Content-based annotation
Ontological T-Box (Taxonomies)
Episodical database / Onomasticon
Semantic annotations
Situation Graph Tree
SGT Traversal
Human motion model
Conceptual scenario model
Reasoning engine (F-Limette)
INTERPRETER MODULE
Spatiotemporal tracking data

Figure 4.7: Scheme of the interpreter module. This module (i) conceptualizes new motion data, (ii) infers new facts from this data using prior models, and (iii) contextualizes the facts to interpret situations.

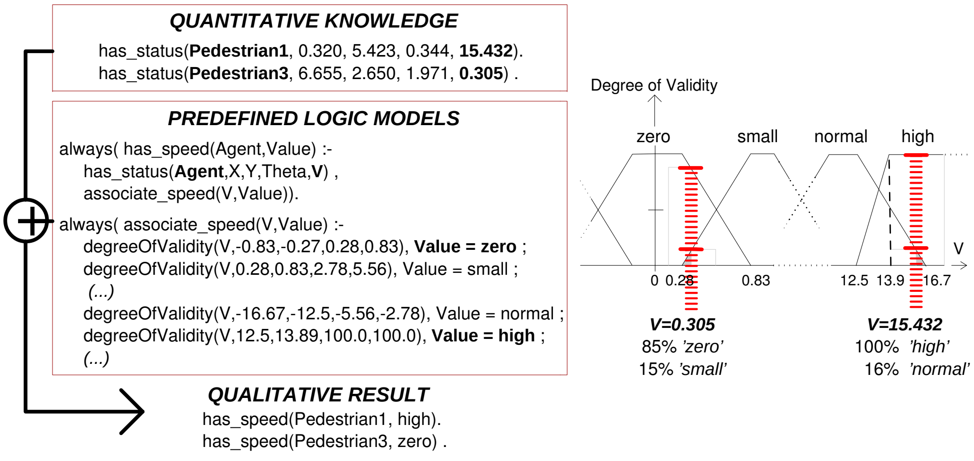
Consequently, a series of high-level interpretations relate and describe at a high-level the relations among entities, objects, and locations over time. For example, Fig 4.10 shows a situation graph that evaluates whether an object has been left, abandoned, or stolen by someone.

An important advantage of our proposal is that the high-level predicates that we use as interpretations of events are actually instantiating ontological relationships. Each generated predicate is mapped to an event from the Event-TBox, which is defined by a series of constraints with entities. These events, constraints, and entities from the TBox are instantiated by facts, constraint instances, and entity instances from the A-Box, respectively, as shown in Fig. 4.11. This way, the tracked entities in a scene are identified as participants of the events, and it is possible for us to easily



(a) (b)

Figure 4.8: (b) Spatio-conceptual models associated to (a) indoor and outdoor testing scenarios.



QUANTITATIVE KNOWLEDGE

```
has_status(Pedestrian1, 0.320, 5.423, 0.344, 15.432).
has_status(Pedestrian3, 6.655, 2.650, 1.971, 0.305) .
```

PREDEFINED LOGIC MODELS

```
always( has_speed(Agent,Value) :-
    has_status(Agent,X,Y,Theta,V) ,
    associate_speed(V,Value)).
always( associate_speed(V,Value) :-
    degreeOfValidity(V,-0.83,-0.27,0.28,0.83), Value = zero ;
    degreeOfValidity(V,0.28,0.83,2.78,5.56), Value = small ;
     (...)
    degreeOfValidity(V,-16.67,-12.5,-5.56,-2.78), Value = normal ;
    degreeOfValidity(V,12.5,13.89,100.0,100.0), Value = high ;
    (...)
```

QUALITATIVE RESULT

```
has_speed(Pedestrian1, high).
has_speed(Pedestrian3, zero) .
```

(a) **(b)**

Figure 4.9: Conversion from quantitative to qualitative values. (a) Input has_status predicates contain tracking data, which is associated to conceptual descriptions. (b) FMTL includes fuzzy mechanisms accepting more than one single interpretation, since it confers *degrees of validity* to values on uncertain ranges.

store a structured registry of their developments over time. Moreover, given that the information is stored in an ontology, this information can be derived to new forms of implicit knowledge through automated inference, thus obtaining a more complete

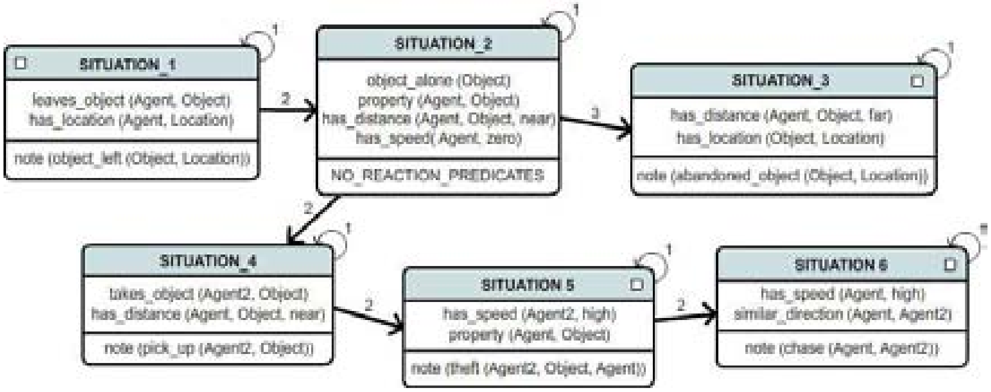

SITUATION_1
leaves_object (Agent, Object)
has_location (Agent, Location)
note (object_left (Object, Location))
SITUATION_2
object_alone (Object)
property (Agent, Object)
has_distance (Agent, Object, near)
has_speed(Agent, zero)
NO_REACTION_PREDICATES
SITUATION_3
has_distance (Agent, Object, far)
has_location (Object, Location)
note (abandoned_object (Object, Location))
SITUATION_4
takes_object (Agent2, Object)
has_distance (Agent, Object, near)
note (pick_up (Agent2, Object))
SITUATION 5
has_speed (Agent2, high)
property (Agent, Object)
note (theft (Agent2, Object, Agent))
SITUATION 6
has_speed (Agent, high)
similar_direction (Agent, Agent2)
note (chase (Agent, Agent2))
1
2
3

Figure 4.10: This situation graph detects that an object has been left by the pedestrian who owns it. The set of conditions are FMTL predicates, the reaction predicate is a `note` command which generates a high-level semantic tag.

registry of occurrences.

Such a structured database is especially useful to let external users interact with the system. We have identified two direct applications to this framework:

1. *Automatic recognition and indexing of video events.* Users have at their disposal a series of semantic annotations over time, which can be filtered by nature, and which partition the video sequence in connected meaningful episodes.

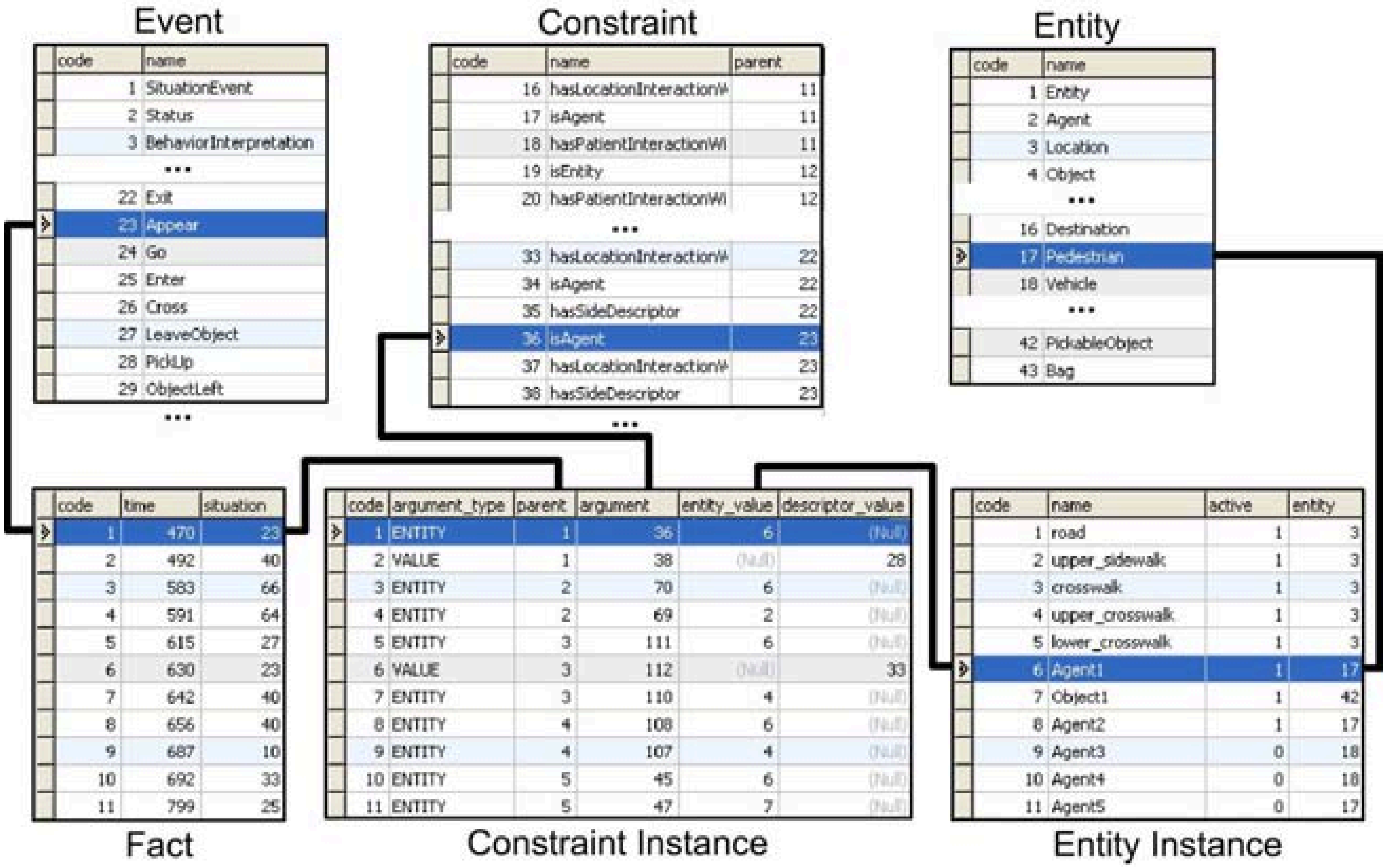
2. *Content-based video retrieval.* Having a registry of developments is useful for users who want to retrieve past information, or search for registered occurrences.

These two applications will be enhanced in the next chapter of this thesis, by providing Natural Language interfaces to ease the communication with external users.

4.6 Application 1: Event annotation

Figs. 4.12 and 4.13 show current experimental results for the annotation of events, in which a collection of high-level predicates have been successfully generated for sequences recorded in outdoor and indoor surveilled scenarios, respectively.[2] The collection of high-level predicates describe interactions among the involved entities, viz. agents, objects, and locations, and also interpretations of behaviors in the case of complex occurrences. Some captures showing the results after tracking processes have been provided, too, for illustration purposes. The number of frame appears in front of each produced annotation, and also in the upper-right corner of each capture. Detections of new agents within the scene have been marked in blue, annotations for activating predefined alerts have been emphasized in red.

[2] The sequences presented are part of the dataset recorded for the HERMES Project (IST 027110, http://www.hermes-project.eu), which has been made available to the scientific community.



Event

code	name
1	SituationEvent
2	Status
3	BehaviorInterpretation
	•••
22	Exit
23	Appear
24	Go
25	Enter
26	Cross
27	LeaveObject
28	PickUp
29	ObjectLeft
	•••

Constraint

code	name	parent
16	hasLocationInteractionW	11
17	isAgent	11
18	hasPatientInteractionWi	11
19	isEntity	12
20	hasPatientInteractionWi	12
	•••	
33	hasLocationInteractionW	22
34	isAgent	22
35	hasSideDescriptor	22
36	isAgent	23
37	hasLocationInteractionW	23
38	hasSideDescriptor	23
	•••	

Entity

code	name
1	Entity
2	Agent
3	Location
4	Object
	•••
16	Destination
17	Pedestrian
18	Vehicle
	•••
42	PickableObject
43	Bag

code	time	situation
1	470	23
2	492	40
3	583	66
4	591	64
5	615	27
6	630	23
7	642	40
8	656	40
9	687	10
10	692	33
11	799	25

Fact

code	argument_type	parent	argument	entity_value	descriptor_value
1	ENTITY	1	36	6	(Null)
2	VALUE	1	38	(Null)	28
3	ENTITY	2	70	6	(Null)
4	ENTITY	2	69	2	(Null)
5	ENTITY	3	111	6	(Null)
6	VALUE	3	112	(Null)	33
7	ENTITY	3	110	4	(Null)
8	ENTITY	4	108	6	(Null)
9	ENTITY	4	107	4	(Null)
10	ENTITY	5	45	6	(Null)
11	ENTITY	5	47	7	(Null)

Constraint Instance

code	name	active	entity
1	road	1	3
2	upper_sidewalk	1	3
3	crosswalk	1	3
4	upper_crosswalk	1	3
5	lower_crosswalk	1	3
6	Agent1	1	17
7	Object1	1	42
8	Agent2	1	17
9	Agent3	0	18
10	Agent4	0	18
11	Agent5	0	17

Entity Instance

Figure 4.11: Detail of the structured relations between concepts and instances in the factual database: upper tables contain TBox concepts (events, constraints, and entities), and lower ones show their A-Box instances.

The outdoor scene was recorded with 4 static cameras and 1 active camera. The video sequence contains 1611 frames (107 seconds) of 720×576 pixels, in which pedestrians, pickable objects, and vehicular traffic are involved and interrelated in a pedestrian crossing scenario. A total of 3 persons, 2 bags, and 2 cars appear on it. The events detected within the scene range from simple agents entering and leaving the scenario to interpretations of behaviors, such as objects being abandoned in the scene, a danger of runover between a vehicle and two pedestrians, or a chasing scene between two pedestrians.

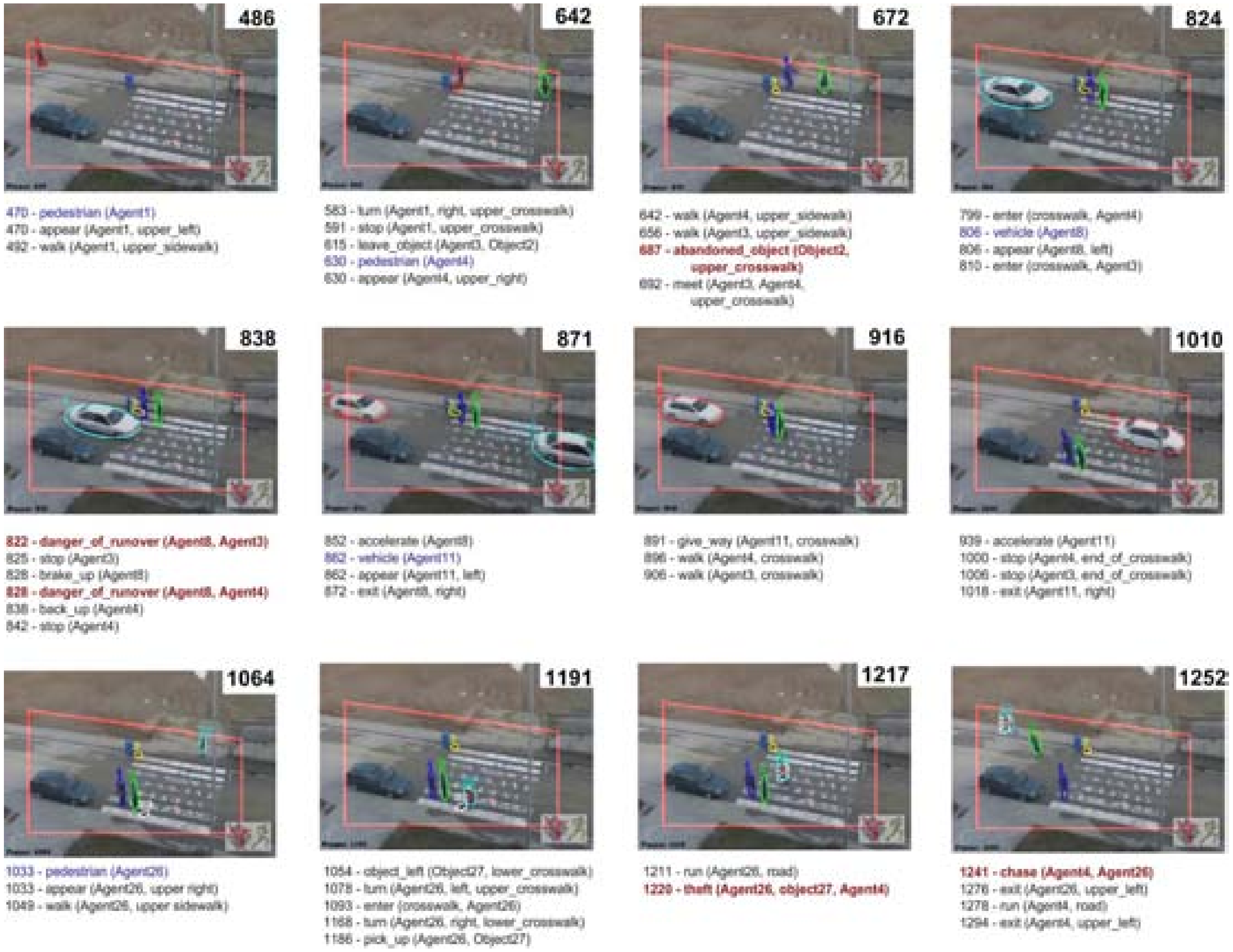
The indoor scene was also recorded with 4 static cameras and 1 active camera. The scene contains 2005 frames (134 seconds) of 1392×1040 pixels, in which 3 pedestrians and 2 objects are shown interrelating among them and with the elements of a cafeteria, e.g. a vending machine, chairs, and tables. The events instantiated in this case include again agents appearing and leaving, changes of position among the different regions of the scenario, sit-down and stand-up actions, and behavior interpretations such as abandoned objects (in this case this is deduced once the owner leaves the surveilled area), the interaction with a vending machine, and violent behaviors such as kicking or punching elements of the scenario.

The proposed approach for situation analysis is capable of carrying and managing confidence levels, obtained at the conceptual stage in form of degrees of validity for the FMTL predicates. Nevertheless, the current implementation relies on the assertion of those predicates associated with the highest confidence values, in order to avoid a combinatorial explosion of solutions. As a consequence, only one high-level predicates is produced by the SGT at each frame, which permits to associate each predicate with an interval of validity.

Part of the evaluation has been accomplished by means of NL input queries over the two presented scenes. At this regard, a list of 110 possibly interesting NL questions or commands to formulate have been proposed by a group of 30 persons from different sources in 5 countries. The current capabilities have been restricted to those user inputs representable by the set of goal queries described in the previous section. Complex input queries such as those related to pragmatic content, e.g. "*Why has the second person come back?*" or "*How is the last pedestrian crossing the road?*", cannot be answered by the system at present and will be tackled in further steps.

Other evaluation results for the current implementation have highlighted that an increment of complexity especially affects two tasks in the high-level architecture: the evaluation of FMTL predicates by the inference engine and the access to the ontology. An increment of length for the recorded sequences results in an exponential growing of the instantiated elements in the conceptual database, and as a consequence a higher increment in the computational time for the SGT traversal. These results encourage the use of heuristic methods to solve these difficulties.

When an alarm is missed from the Vision levels, the hierarchical structure of the SGT simply does not instantiate a situation, since one of its required state conditions is not accomplished. If the rest of information does not allow to reach a certain level of specialization for a situation, then its parent situation will be asserted. Otherwise, a general situation will be asserted due to the lack of information. Thus, the more exhaustively we define the hierarchy of a SGT, the more robust will be the system in front of missing information, but the more expensive it will be the cost in terms of



470 - pedestrian (Agent1)
470 - appear (Agent1, upper_left)
492 - walk (Agent1, upper_sidewalk)

583 - turn (Agent1, right, upper_crosswalk)
591 - stop (Agent1, upper_crosswalk)
615 - leave_object (Agent3, Object2)
630 - pedestrian (Agent4)
630 - appear (Agent4, upper_right)

642 - walk (Agent4, upper_sidewalk)
656 - walk (Agent3, upper_sidewalk)
687 - abandoned_object (Object2, upper_crosswalk)
692 - meet (Agent3, Agent4, upper_crosswalk)

799 - enter (crosswalk, Agent4)
806 - vehicle (Agent8)
806 - appear (Agent8, left)
810 - enter (crosswalk, Agent3)

822 - danger_of_runover (Agent8, Agent3)
825 - stop (Agent3)
828 - brake_up (Agent8)
828 - danger_of_runover (Agent8, Agent4)
838 - back_up (Agent4)
842 - stop (Agent4)

852 - accelerate (Agent8)
862 - vehicle (Agent11)
862 - appear (Agent11, left)
872 - exit (Agent8, right)

891 - give_way (Agent11, crosswalk)
896 - walk (Agent4, crosswalk)
906 - walk (Agent3, crosswalk)

939 - accelerate (Agent11)
1000 - stop (Agent4, end_of_crosswalk)
1006 - stop (Agent3, end_of_crosswalk)
1018 - exit (Agent11, right)

1033 - pedestrian (Agent26)
1033 - appear (Agent26, upper right)
1049 - walk (Agent26, upper sidewalk)

1054 - object_left (Object27, lower_crosswalk)
1078 - turn (Agent26, left, upper_crosswalk)
1093 - enter (crosswalk, Agent26)
1168 - turn (Agent26, right, lower_crosswalk)
1186 - pick_up (Agent26, Object27)

1211 - run (Agent26, road)
1220 - theft (Agent26, object27, Agent4)

1241 - chase (Agent4, Agent26)
1276 - exit (Agent26, upper_left)
1278 - run (Agent4, road)
1294 - exit (Agent4, upper_left)

Figure 4.12: Set of semantic annotations produced for the outdoor scene, which have been automatically generated for the fragment of recording comprised between frames 450 and 1301.

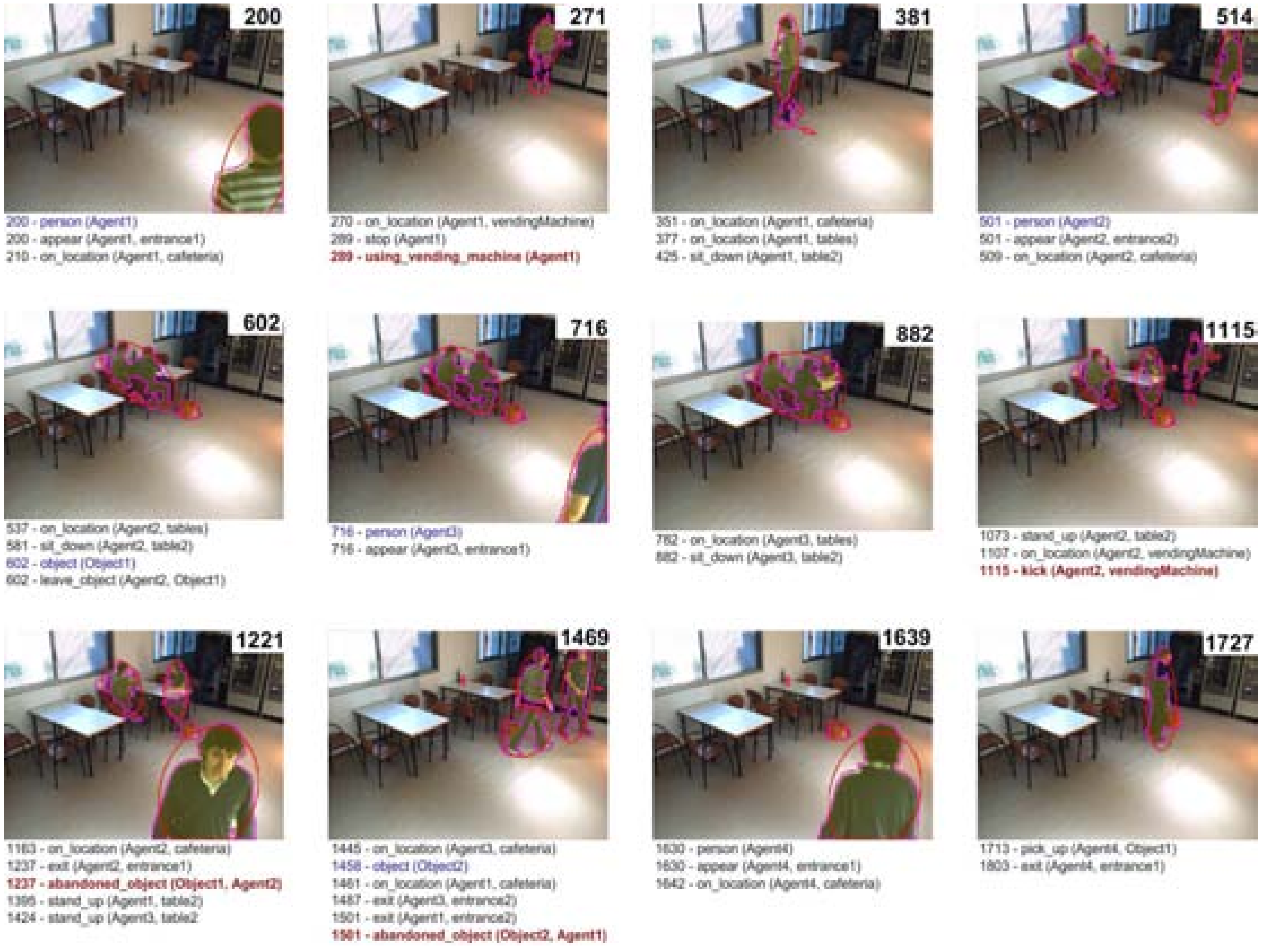

200
200 - person (Agent1)
200 - appear (Agent1, entrance1)
210 - on_location (Agent1, cafeteria)
271
270 - on_location (Agent1, vendingMachine)
289 - stop (Agent1)
289 - using_vending_machine (Agent1)
381
351 - on_location (Agent1, cafeteria)
377 - on_location (Agent1, tables)
425 - sit_down (Agent1, table2)
514
501 - person (Agent2)
501 - appear (Agent2, entrance2)
509 - on_location (Agent2, cafeteria)
602
537 - on_location (Agent2, tables)
581 - sit_down (Agent2, table2)
602 - object (Object1)
602 - leave_object (Agent2, Object1)
716
716 - person (Agent3)
716 - appear (Agent3, entrance1)
882
792 - on_location (Agent3, tables)
882 - sit_down (Agent3, table2)
1115
1073 - stand_up (Agent2, table2)
1107 - on_location (Agent2, vendingMachine)
1115 - kick (Agent2, vendingMachine)
1221
1163 - on_location (Agent2, cafeteria)
1237 - exit (Agent2, entrance1)
1237 - abandoned_object (Object1, Agent2)
1395 - stand_up (Agent1, table2)
1424 - stand_up (Agent3, table2
1469
1445 - on_location (Agent3, cafeteria)
1458 - object (Object2)
1461 - on_location (Agent1, cafeteria)
1487 - exit (Agent3, entrance2)
1501 - exit (Agent1, entrance2)
1501 - abandoned_object (Object2, Agent1)
1639
1630 - person (Agent4)
1630 - appear (Agent4, entrance1)
1642 - on_location (Agent4, cafeteria)
1727
1713 - pick_up (Agent4, Object1)
1803 - exit (Agent4, entrance1)

Figure 4.13: Set of semantic annotations produced for the indoor scene, which have been automatically generated for the fragment of recording comprised between frames 150 and 1839.

computation.

A similar consideration has to be done regarding false alarms: the SGT will instantiate a wrong situation only when the false information agrees with the sequence of admissible states defined in the tree by means of the prediction edges. This way, the robustness of the situational analysis is given by the SGT based on both the temporal and specialization criteria. The generation of incorrect information depends of both the sensory gap (bad information provided by the vision acquisition systems) and the semantic gap (incorrectness or incompleteness of the models at high level).

These experimental results for the situational analysis have been obtained using the F-Limette[3] inference engine for fuzzy metric-temporal horn logic and the SGTEditor[4] graphical editor for SGTs. On the other hand, the implementation of the ontology and the query system have been developed using the Protégé[5] ontology editor and the Jena[6] Semantic Web Framework.

For evaluation purposes, we have compared the automatic annotations given by the system with those given by a significant amount of population. Different image sequences from the same domain have been used to train the system and to test its performance.

The ground truth annotation of events was accomplished using 3 different image sequences, 2 outdoor and 1 indoor. The first outdoor sequence (2250 frames@25fps, 640×480 pixels) shows the entrance of a public building, where pedestrians come in and out and interact with some cars and motorbikes on their way. The second outdoor sequence (600 frames@15fps, 1256×860 pixels) is a crosswalk scenario, in which 4 pedestrians enter a crosswalk in different manners, in the presence of vehicular traffic. The indoor training video (1575 frames@15fps, 1256×860 pixels) contains specific events like leaving bags, greeting a person, taking objects from someone else, sitting down, or kicking a vending machine.

Two scenes from the same domain were recorded for tests, one in a traffic scenario and the other one in a cafeteria, see Fig. 4.8. These test scenes share similar events than the ones found in the test sequences, in completely different scenarios. The outdoor scene contains 1611 frames@15fps of 720×576 pixels, in which pedestrians, pickable objects, and vehicular traffic interact in a pedestrian crossing. The indoor scene contains 2005 frames@15fps of 1392×1040 pixels, in which people and objects interact among them and with the elements of a cafeteria, viz. a vending machine, chairs, and tables. Both sequences show complex events like abandoned objects, thefts, chases, or vandalism. These sequences have been automatically analyzed and indexed by the proposed system.[7]

The asserted events for every detected target have been stored in a SQL relational database to enable data retrieval. Every asserted event points to a temporal interval of validity in the sequence, and relates the involved target to its contextual blanket. The collection of video annotations describe interactions among the involved entities, and also interactions and interpretations of complex occurrences.

[3] http://cogvisys.iaks.uni-karlsruhe.de/Vid-Text/f_limette/index.html

[4] http://cogvisys.iaks.uni-karlsruhe.de/Vid-Text/sgt_editor/index.html

[5] http://protege.stanford.edu/

[6] http://jena.sourceforge.net/

[7] The sequences used in these experiments can be found at http://iselab.cvc.uab.es/ tools-and-resources.

Entity type ($\mathcal{T}$)	Instance ($\mathcal{A}$)
Pedestrian	*ped2*
Vehicle	*veh1*
Location	*sidewalk*
Object	*obj1*
Descriptor	*fast*

Event type ($\mathcal{T}$)	Indexed fact ($\mathcal{A}$)
Spatiotemporal	*walk (ped2, fast)*
Interaction	*appear (ped2, sidewalk)*
Interaction	*pick_up (ped2, obj1)*
Interpretation	*theft (ped2, ped1, obj1)*
Interpretation	*danger_of_runover (veh1, ped2)*

Table 4.7

Possible instances of entities (left) used in event indexes (right). For a *theft* to be indexed, *ped2*, *ped1*, and *obj1* must accomplish a certain semantic context.

4.7 Application 2: Content-based video retrieval

Regarding content-based video retrieval, we tested how many and which kind of queries provided by a set of volunteers were understood and correctly answered by the system. The details about NL components present in this experiment have been purposely omitted in this section, since at this point we are interested in the system's management of semantics. Next section describes thoroughly the modules used to enable the conversion from semantic predicates to linguistic expressions.

Examples of content-based video retrieval are presented in Table 4.8, which retrieve episodes of sequences containing certain events or entities. More complex queries are possible, e.g. querying for chases after thefts, objects owned by different persons, or scenes in which a number of agents were seen at a certain location. As for the NL queries, acceptable propositions also restrict to the domain imposed by the ontology. This way, users were enabled to ask for any modeled event involving any of the entities, which is related to any semantic zone in the scenario, and happens at any point or interval of time. These are some examples of the most repeated types of user queries that have been accepted by the NL module:

- *Show me pedestrians meeting between frames 300 and 1200.*
- *How many people has picked up bags?*
- *Have you seen any pedestrian running by the road after a theft?*
- *List all vehicles before frame 600.*

Similar concepts are automatically linked using the metrics over WordNet, such as *pedestrians–people*. In the experiments, subjects usually restricted to simpler queries. The difficult queries were usually too generic or stepped out of the domain, with sentences such as "*How is this person dressing?*" or "*Does it rain?*", in which case the concepts found could not be linked to the factual database. Out of the total number of queries asked that belonged to the domain, a 91% of them led to proper understanding by the system. Most of the non-understood questions were those starting with *why* or *how*, types that usually result less objective to answer.

These results have been compared to the validation data set provided by a second group of subjects. Fig. 4.14 shows the number of events agreed by a certain percentage

Entity ID: **Agent5**
Interval: 1200–1250
Sequence: Outdoor-1

Interval	Event	Arguments

1186–1202	pick_up	is_agent(**Agent5**)
		has_object_interaction_with(Object2)

1186–1276	carry_object	is_agent(**Agent5**)
		has_object_interaction_with(Object2)

1211–1219	run	is_agent(**Agent5**)
		has_location_interaction_with(Road)

1220–1240	theft	is_agent(**Agent5**)
		has_patient_interaction_with(Agent1)
		has_object_interaction_with(Object2)
		has_property(Malicious)

1241–1275	chase	is_agent(Agent1)
		has_patient_interaction(**Agent5**)

Entity ID: **Object1**
Interval: 550–1250
Sequence: Indoor-2

Interval	Event	Arguments

501–601	carry_object	is_agent(Agent2)
		has_object_interaction(**Object1**)

602–1236	leave_object	is_agent(Agent2)
		has_object_interaction(**Object1**)
		has_location_interaction(Hall)

1237–1712	abandoned_object	is_patient(Agent2)
		has_object_interaction(**Object1**)
		has_property(Malicious)

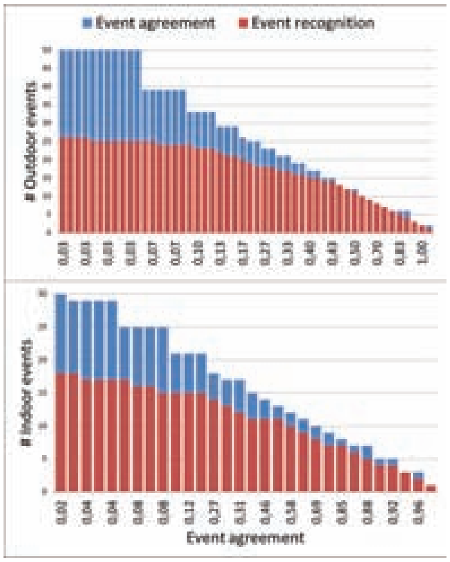
Table 4.8

Examples of retrieval of episodic events when querying for a given entity.

of the population (*event agreement*), and the events out of that set correctly identified by the system (*agreed event recognition*, or simply, *event recognition*). Fig. 4.15 presents the percentage of events correctly recognized. As we can see, for sets of events agreed by above 50% of the population, the system recognizes all of them in the outdoor scenario and 85% of them in the indoor one. On the other hand, if we consider the set of events identified by more than 90% of the subjects, a recognition rate of more than 90% is achieved in both scenarios.

Some examples of non-recognized annotations are *ignore_object*, *be_upset*, *be_hesitant*, *talk*, *realize_about_someone*, or *shake_hands*, among others, which mostly happened in indoor sequences. All undetected events were shared by less than 20% of the population, given the subjectivity of the interpretation, except for *talk* and *shake_hands*. In these two cases, the semantic framework facilitates retrieving non-modeled events by searching for similar concepts, e.g. *meet* or *interact*.

The reason of the different performance between indoor and outdoor scenes is that although indoor image sequences permit a reduced viewpoint and incorporate less events, the events detected show a higher semantics, such as body gestures, facial expressions, and subtler interactions between agents, which require more knowledge than that one obtained solely from trajectory data.



Outdoor indexing	
Event agreement	Event recognition
0.03	0.52 (26/50)
0.10	0.73 (24/33)
0.20	0.76 (19/25)
0.33	0.81 (17/21)
0.40	0.88 (15/17)
0.50	1.00 (12/12)
0.60	1.00 (10/10)
0.70	1.00 (8/8)
0.80	1.00 (7/7)
0.90	1.00 (3/3)
1.00	1.00 (2/2)

Indoor indexing	
Event agreement	Event recognition
0.03	0.62 (18/29)
0.11	0.71 (15/21)
0.25	0.78 (14/18)
0.31	0.76 (13/17)
0.42	0.73 (11/15)
0.50	0.85 (12/13)
0.61	0.82 (9/11)
0.69	0.80 (8/10)
0.80	0.78 (7/9)
0.92	0.80 (4/5)
1.00	1.00 (1/1)

Figure 4.14: Correctly indexed events. Left graphic: horizontal axis shows the percentage of people agreeing with a set of events; vertical axis reports the total of events in this set, and the number out from them that were recognized. Right table: numeric details.

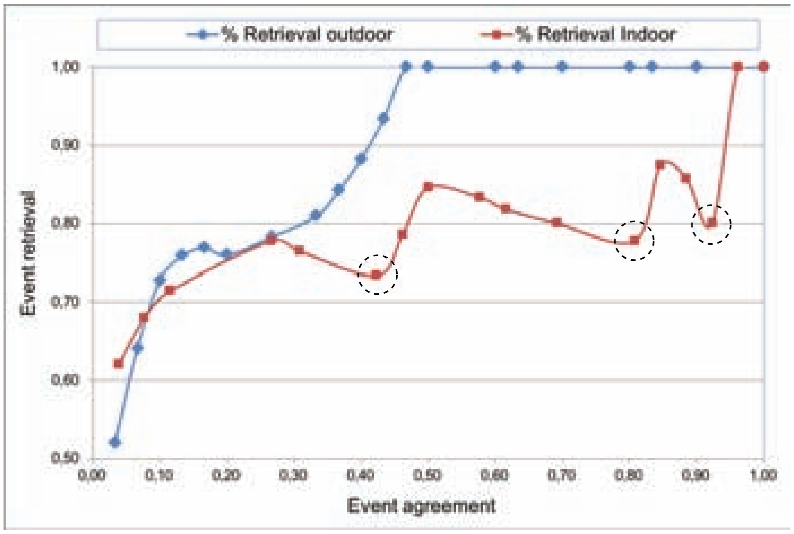


Figure 4.15: Percentage of retrieval. Failures in indoor sequences are mainly due to unhandled recognition of expressions and gestures by the vision algorithms. Highlighted minima correspond to *be_upset*, *shake_hands*, and *talk* (left to right).

4.8 Extension to *Fuzzy Constraint Satisfaction*

An excessive determinism could be argued as a critical issue of our interpretation module. A large variety of probabilistic detectors and classifiers contribute to the current state-of-the-art on action recognition, and it would be sensible to benefit from such robust outputs. However, categorizing them into *true/false* predicates would discard valuable statistical information. The best strategy would be to preserve both a *logical* and a *quantitative* form, and use them conveniently. Recent trends on fuzzy logic and DL can help us to preserve uncertainty in the logical inferences, while additionally incorporating the task into our ontological framework. Our most recent steps on event recognition follow this direction.

Several techniques allow us to infer mid-level concepts using motion cues. The problem of abandoned objects, for example, is usually solved using background substraction or blob dynamics –e.g., a blob splits into two, and one of them remains still until being absorbed by the background [43]–. In addition, trajectories through regions of interest suggest particular behaviors given an adequate scene prior. For instance, if we model a `paying` event as a person interacting with an automatic cashier before walking back to a car, we need to assert (i) that the blob moving towards the cashier is a person, and that (ii) it is actually paying. For complex atomic actions like these, more sophisticated techniques based on statistical learning are required [113, 31, 73]. We propose to define fuzzy rules that incorporate multiple sources of uncertainty, and reason about them in order to assign confidences to each defined event. A fuzzy reasoner based on DL, `fuzzyDL` [21], has been used as a framework to define a knowledge base of spatiotemporal occurrences, and eventually perform the reasoning.

A suitable formalization of our problem is posed in terms of a *Fuzzy Constraint Satisfaction* (FCS) problem [110]. It is formally defined as follows: let us consider a set of fuzzy variables $V = \{V_1, \ldots, V_m\}$ over domains $D_1 \ldots, D_m$, respectively. For instance, we define crisp domains in which membership functions assign a so-called Degree of Satisfaction $DoS \in [0, 1]$. Let us also consider a set of constraints $C = \{C_1, \ldots, C_n\}$, each one ranging over a subset of V. The goal is to find an assignment of values $(d_1, \ldots, d_m) \in D_1 \times \cdots \times D_m$ such that $C_1, \ldots, C_n$ are satisfied, or in other words, to obtain a variable assignment that is optimal with respect to the DoS of V and C. To find the optimal assignment, a joint DoS of each variable V_i is defined as

$$DoS(V_i) := \frac{1}{w+1}\left(\frac{1}{|C_i^+ + C_i^-|}\left(\sum_{c \in C_i^+} c + |C_i^-|\right) + w\mu_i(l_i)\right) \tag{4.3}$$

where w is the weight for that particular variable, C_i^+ is the DoS of a variable assignment for each fully instantiated constraint, and C_i^- is the overestimated DoS for each partially instantiated constraint.

The `meet` event

For simplicity, we have chosen to model a *meet* event between two individuals, since (i) it is intrinsically fuzzy, i.e., it is easier to explain using vague terms rather than strict formulae, and (ii) it is interesting enough to have been extensively tackled by the research community [96, 3, 103, 100]. Nonetheless, the work described can be extrapolated to different scenarios, such as outdoor surveillance in parking lots or toll barriers in highways. In these cases, for instance, actions that could be interesting to identify may include collisions or scratching between cars in a parking lot, or proper/improper payments via cashier after parking a car.

To detect a *meet* event we require tools for trajectory analysis, action recognition, and a definition of some space-time constraints. In [14], we find a new method for classification of human actions that relies on an appropriate quantization process, dealing with the ambiguity of the traditional codebook model. The analysis of trajectories is granted by means of a simple blob detector from the OpenCV library[8]. Finally, a simple fuzzy rule measures quantitatively the confidence on two persons meeting, by inferring the concepts `closeness` and `previous_closeness` upon the metric distances between two subjects at the current moment and 15 frames before. For this example we have considered 3 membership functions: `far`, `medium`, and `close`. The estimated metric distance d is normalized into a range $[0, 1]$ using the mapping $\tilde{d} = \exp(-\lambda d)$, where we assume medium distance as $d = 3m$ ($\lambda = \frac{1}{3} \ln 2$). Finally, the following rule has been modeled in order to detect a `meet` occurrence:

$$\exists \texttt{previous\_closeness}(\texttt{medium}) \wedge \exists \texttt{closeness}(\texttt{close}) \Rightarrow \texttt{meet}(\texttt{meeting}) \qquad (4.4)$$

where the existential operator is defined as the conjunction of a relation and an unary concept as follows: $\exists R(C_1) \equiv \sup_{y \in \Delta^I} R^I(x, y) \oplus C_1^I(y)$. The quantitative estimation of a meeting comes as the defuzzification of the `meet` concept using the largest of its interpretation maxima. Fig. 4.16 shows an example of defuzzification of the concept `meet`, extracted from the sequence used for evaluation.

Implementation and results

A short video sequence of 235 frames has been recorded, in which two pedestrians approach to each other at a normal pace, one of them suddenly does a gesture towards the second one and runs away, and the second one chases after him. The sequence has been tracked, and the generated spatiotemporal data has been reasoned using the described framework.

Fig 4.17 depicts the obtained results. The top row shows 6 snapshots taken at different frames of the sequence, and below there is a ground-plane reconstruction of the instantaneous position of each target. The tracker has lost target 1, renaming it as target 4; in addition, an extra blob –the head of pedestrian 2– has been wrongly detected as an additional target 3. The bottom row depicts the numerical (defuzzified) confidence on a fuzzy meeting, where green zones stand for intervals with asserted `meet` predicates.

The modeled rules interpret correctly the dynamic interactions between blobs 1 and 2 –first approaching, second move–, although this setup is very sensitive to the errors of the tracker. In case of lack of precision, the blobs would be ill-projected to the ground plane, and projectivity would amplify the initial errors. Nonetheless, the

[8] http://sourceforge.net/projects/opencvlibrary/

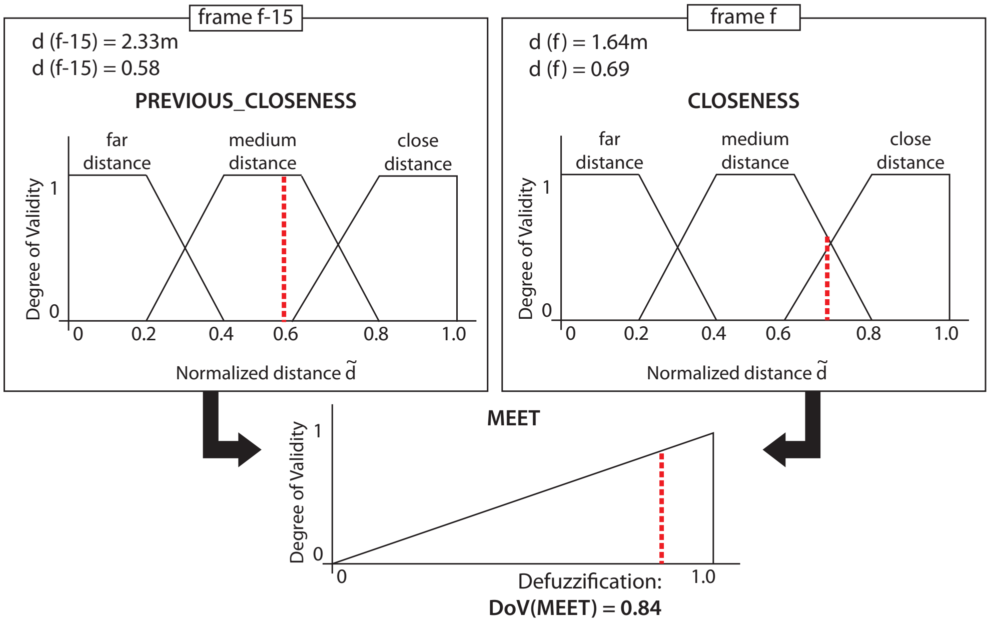
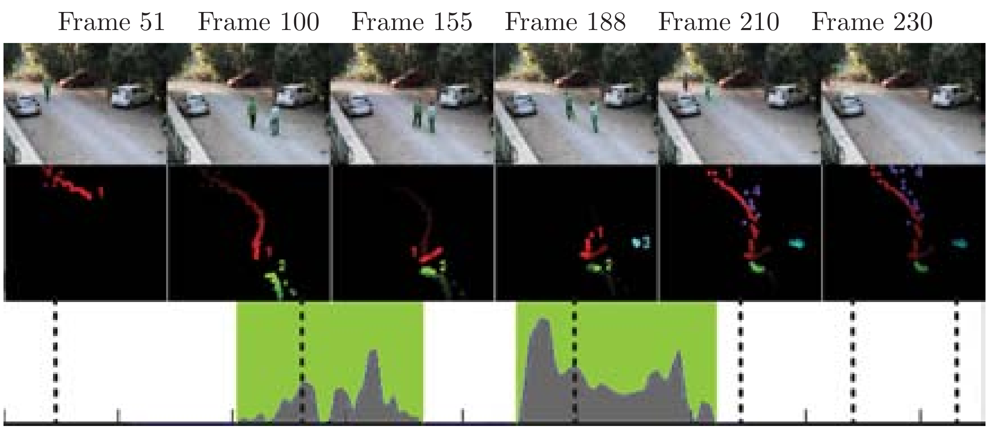


Figure 4.16: Detection of a meeting. The estimated distances are fuzzy-conceptualized, and the reasoner defuzzifies a `meet` value using Eq. 4.4.



Defuzzified `meet(Pedestrian1, Pedestrian2)` event, with markers for the 6 frames

Figure 4.17: Preliminary results assessing the confidence on a 2-person *meet* event (bottom row). Middle row shows last rectified ground-plane positions observed.

utility of this framework is promising. The use of better detectors and trackers would probably facilitate the modeling of more complex events.

4.9 Discussion

State-of-the-art on smart video analysis is heading to the automatic exploitation of semantic context, in order to extract event patterns that permit us a better comprehension of image sequences. Nevertheless, few works assess the suitability and coverage of the selection of semantic events to model, and most of them are restricted to very specific scenarios, thus questioning the generalization capability of the methods used. In addition, these events should also be suited for end-user interfacing of video contents, something difficult to achieve by only using bottom-up procedures.

Our methodology contributes to these three challenges. First, it copes with the ambiguous and sometimes incorrect interpretations done by experts while building conceptual models. The ontology and the rest of the knowledge bases are modeled in a top-down manner from users' textual evidence, constituting a separate identifiable part of the design. The technique chooses the most suited event concepts from different scenarios, merging them into single models –ontology, SGT–, and thus enabling generalization to different scenarios in the domain. Finally, since the ontology has been built from linguistic corpora, it provides straightforward connection to NL interfaces like those shown for video description and retrieval, allowing end-users to access meaningful video content flexibly by means of NL descriptions and dialogue-based instructions.

The Event-TBox provides the space of validity of possible semantic video indexes in a domain. The ontological constraints applied to the terminologies fix the validity of situations to detect; this way, mechanisms for prediction based on restrained behavioral models can be developed. High-level predicates have been chosen as the central semantic elements of the cognitive vision system, for them being highly expressive, language-independent, and suitable for a neutral framework between vision and linguistics. The most basic events are defined by generic, domain-independent human motion models.

An SGT acts as an actual content classifier, which semantically characterizes the temporal intervals of video sequences: the resulting predicates can be identified as high-level semantic indexes, which facilitate further applications such as search/retrieval/browsing engines. The modular dimension of this framework provides multimodality: arbitrary modules providing new types of data can be directly incorporated, and as long as this data is made available in form of conceptual facts, it easily integrates into the situation analysis. In addition, the presented approach directly benefits from the automatic learning of semantic regions described in the previous chapter.

To consolidate the interpretation process, next steps should enhance SGTs in order to let them hold multiple hypotheses as probable interpretations during the traversals. Future work should also be directed to study extensions of the proposed framework to the challenging domain of movie and media analysis. To this end, current behavioral models need to be enhanced by modules that enable the system to recognize body postures and facial expressions, taking advantage of the high resolution typically found in video data from these domains. The behavior of crowds and large groups of agents has not been analyzed yet, and will as well be included as future work.

Resum

L'estat de l'art de l'anàlisi intel·ligent de vídeo s'està dirigint cap a l'explotació automàtica de context semàntic, de cara a extreure patrons d'activitat que ens permetin arribar a comprendre millor les seqüències d'imatges. Tot i això, no hi ha massa recerca que proposi formes adequades de selecció de conceptes per a modelar activitats, i la majoria es restringeix a escenaris particulars; per tant, es qüestiona la capacitat de generalització dels mètodes resultants. A més, seria desitjable que el modelat d'events pogués beneficiar directament les interfícies avançades d'usuari per a l'exploració de continguts de vídeo, cosa que resulta difícil emprant només inferència ascendent.

La proposta té tres contribucions principals. Primerament, es suggereix una sistemàtica per al modelat conceptual que permet evitar interpretacions incorrectes o ambígües per part d'experts. L'ontologia i altres bases de coneixement es modelen descendentment en base a l'evidència textual proporcionada pels usuaris. D'aquesta forma, la tècnica escull els events més adequats per a diversos escenaris, combinant-los en models únics com els SGT i l'ontologia, i així permetent generalitzar el reconeixement d'activitat al domini d'interès. Finalment, donat que l'ontologia es basa en entrenament lingüístic, es pot connectar directament amb interfícies de llenguatge natural com les emprades per a cerques i descripcions lingüístiques de vídeo, deixant a l'abast dels usuaris l'accés a continguts semàntics mitjançant diàlegs.

La taxonomia d'events restringeix l'espai de validesa dels índexs semàntics en el domini. Les restriccions ontològiques imposades als events fixen un seguit de condicions perquè una situació es detecti; així es poden aconseguir mecanismes de predicció basats en models tancats. S'han escollit els predicats d'alt nivell com elements semàntics central del sistema cognitiu, donat que són altament expressius, independents de la llengua, i adequats com a nivell neutral entre la visió i la lingüística. D'altra banda, els events bàsics es defineixen per models humans genèrics i independents del domini.

Els arbres de grafs de situacions (SGT) són de fet classificadors de contingut que caracteritzen semànticament els intervals temporals dins una seqüència de vídeo. Les interpretacions generades actuen com a índexs semàntics d'alt nivell, faciliten aplicacions de cerca, consulta i navegació de vídeo basat en contingut. El fet que el sistema sigui modular facilita la multimodalitat, donat que s'hi poden incorporar arbitràriament mòduls que proporcionin diferents tipus de dades. Sempre que aquestes vinguin en forma de predicats lògics es poden incloure directament als SGT. A més, els resultats d'aprenentatge automàtic de regions presentats al capítol anterior es poden utilitzar directament per l'anàlisi d'alt nivell descrit aquí.

Finalment, hi ha tot un seguit de tasques que cal millorar o incorporar en el futur. Per exemple, els SGT haurien de permetre raonar amb múltiples hipòtesis concurrentment, cosa que no es permet actualment. També s'ha d'estudiar quines dificultats comportaria passar de la vídeo vigilància a l'anàlisi i indexació automàtica de pe·ícules i continguts multimèdia genèrics. Per fer-ho, caldria incorporar l'anàlisi de postures i expressions facials, donat que en els nous dominis la resolució és típicament molt més alta i aquestes tasques serien possibles. Per últim, encara en el camp de la video vigilància, el sistema hauria de saber analitzar el comportament de multituds, que presenta una sèrie de problemes difícils de resoldre.

Chapter 5

Ontology-based human-computer interaction

"And what is the use of a book", thought Alice,
"without pictures or conversations?"

Alice in Wonderland (1865), by Lewis Carrol

The ability to communicate is innate in a natural cognitive system. There exist several ways to reach this goal artificially, although Natural Language is usually taken as a primary choice, being a flexible, unconstrained, and economical tool that is also intrinsic to end-users. This chapter discusses the implementation of linguistic modules to close the communication loop between the system and external users. Additional tasks like the generation of virtual scenes are also implemented and combined, in order to increase the benefits of high-level interfaces for human-machine interaction. This chapter exploits ontological knowledge and user interfaces to narrow many of the gaps –interface, query, model, semantic–.

A fundamental objective of cognitive systems is to achieve effective human-machine interaction. This is useful to enhance the human capability and productivity across a large collection of endeavors related to a definite domain. Some alternatives are available to grant human-machine communication, such as Natural Language and Computer Graphics. In our case, for example, we may think of three particularly interesting types of interaction:

- Generating textual accounts about observed occurrences.
- Understanding textual queries and commands from external users.
- Displaying synthetic videos with virtual elements representing the real scene.

The first task is accomplished by a process of Natural Language text Generation (NLG), and the second by Natural Language text Understanding (NLU). These two first tasks are intrinsically relevant for our goal, for they grant linguistic communication, i.e., the easiest and fastest way for non-expert users to reach out to the system. The third one is attractive as well, for it provides end-users with a simplified representation of the observations, while holding their content. Moreover, this task is also interesting for designers and maintainers: first, it becomes a cheap way to evaluate tracking systems over fully controlled environments, e.g., making scenes complex by gradually incorporating behavioral virtual agents. Secondly, it efficiently compresses hours of video material into a light list of semantic predicates, which virtually recreate the developments anywhere.

This chapter explores ways in which these three contributions can be incorporated to the framework described so far. Next section analyzes some preliminary ideas about NL, especially stating the differences between NLG and NLU tasks. After it, Section 5.3 presents our starting point for embedding linguistic capabilities to the system, the so-called Discourse Representation Theory (DRT). Based on this idea, an initial NLG module is detailed. In Section 5.3, the original module is enhanced to deal with multilingual capabilities and to confer language-independent extensibility. Section 5.4 proposes a NLU module that couples with the ontological resources of the system, thus closing the communication loop. In addition, a module for the generation of synthetic scenes is detailed in Section 5.5. Finally, some experimental results validate the suggested applications.

5.1 Introductory remarks on NL

As introduced above, NL becomes fundamental when discussing the communication with end-users. A natural linguistic communication involves two main capacities: to put words to our thoughts, and to identify thoughts from the words we perceive, and these are the goals that we transfer to the system by means of NLG and NLU. Both tasks are subfields of Natural Language Processing, which in turn can be seen as a subfield of both computer science and cognitive science [106]. NLG focuses on computer systems that automatically produce understandable texts in a natural human language, and NLU studies computer systems that understand these languages. Both are concerned with computational models of language and its use. In general terms, the two processes have the same end points, but opposite directions. One would think then, by looking at the general picture, that there would be many shared processes or resources that could be reused between them.

Nevertheless, the internal operations of these processes hold several differences in character. NLG has been often considered as a process of *choice* –i.e., which is the best way to deliver the information–, whereas, NLU has been best characterized as one of *hypothesis management* –i.e., which response by the system is the user requesting–. In NLG, we have several means available, and must choose the most suitable one to achieve some desired end. In NLU, we must select the most appropriate interpretation out of a multiple set of them, given some input.

Therefore, the strategy adopted to build a NL interface is different for each task, see Fig. 5.1. In NLG we control the set of situations that need to be expressed, and define *one* correct form of expressing that information in a clear and natural way, for

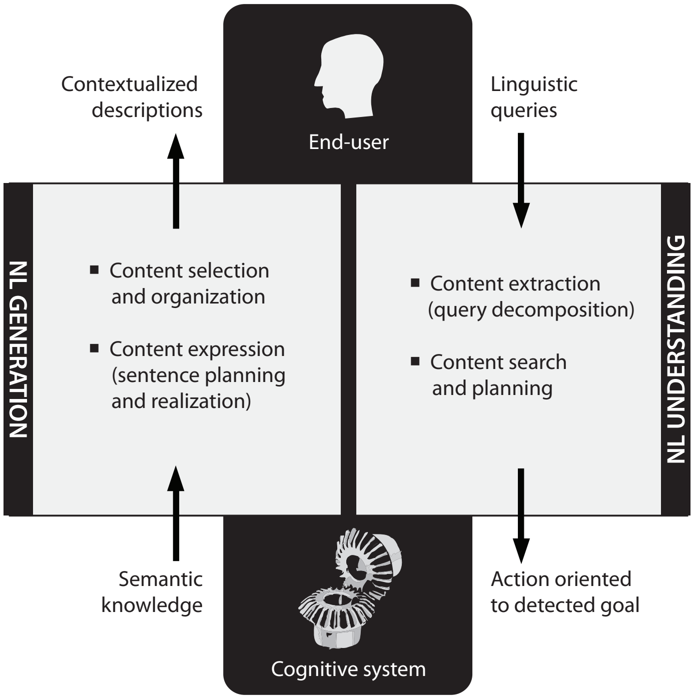

Contextualized descriptions
End-user
Linguistic queries
NL GENERATION
Content selection and organization
Content expression (sentence planning and realization)
Content extraction (query decomposition)
Content search and planning
NL UNDERSTANDING
Semantic knowledge
Cognitive system
Action oriented to detected goal

Figure 5.1: Although NLG and NLU seem to be close related, they involve different problems which require independent strategies.

each language considered. On the other hand, the NLU process provides us with an open number of possible user queries that need to be interpreted; hence, we need to restrict to a set of intentions that we assume a user can show. Following these ideas, the best option is to consider NLG from a closed, deterministic viewpoint and NLU from an open one, since the first one has to do with aforeknown situational models, whereas the second one deals with the unexpected.

From a general perspective, some general guidelines have been considered for a sensible implementation of NLG into our cognitive vision system:

- We must describe situations contained in the implemented behavioral models. In our case, the situations are those defined in a domain ontology and resulting from the behavioral analysis accomplished by SGTs.
- According to the cognitive situatedness/embeddedness property, the behaviors of an agent in a given environment are constrained [134]. Consequently, the system's outputs have been restricted to interpretations of situations uniquely for the defined domain. These interpretations will be expressed linguistically by native speakers of each language, for consistency.
- Such linguistic utterances are built and adapted into the system using rule-based parsing techniques and functional grammars (detailed in Appendix B), which

have been conceived specifically to facilitate multilinguism, extensibility, and effective ontological coupling.

- The final linguistic models are enhanced according to the Prototype Theory from cognitive linguistics, in which the linguistic elements are categorized using sets of semantic features [71]. As explained further on in the text, this approach entails a series of advantages, e.g., the lack of rigidness to formalize linguistic properties, or the interoperability of linguistic knowledge at different stages.

Likewise, we state a series of guidelines for NLU:

- All valid[1] requests that apply to predefined goals in the domain should be detected. Such requests are traditionally classified into questions (queries), commands, and information updates.
- The ontological resources described in the previous chapter are used here to restrict the semantic domain of validity of the possible requests.
- NLU aims to actually understand the intention of a user, so that the system can act according to the hypothetical intention.
- It is valuable to use a probabilistic approach for NLU, given the huge number of possible inputs to express an intention, which in practice cannot be completely controlled. Therefore, the definition of some type of semantic metric is required to assess the most probable interpretation of a request.

Next section describes an implementation for the NLG task, based on a series of recent contributions by different authors related to the field. Subsequently, this first design is enhanced in Section 5.3 to incorporate more functionalities and to link with the existing ontological resources. The enhanced NLG module (ONT-NLG) serves as a basis to derive modules for NLU and generation of synthetic scenes, which eventually cover a wide range of applications targeted to user interaction.

5.2 DRT-based Natural Language Generation (DRS–NLG)

The information to be expressed by the NLG module about a scene is contained in the series of high-level facts stored into the factual database. The main goal for this module consists of selecting a unique form of expressing that knowledge in a clear and natural way, for each of the languages considered. This module is then built from a deterministic point of view, since it deals with aforeknown linguistic realizations.

Reiter and Dale [106] presented a roadmap of the main tasks to be solved regarding NLG. Its proposed model of architecture includes three modules:

- A Document Planner, which produces a specification of the text's content and structure, i.e. *what* has to be communicated, by using both domain knowledge and practical information to be embedded into text.

[1] The validity of an input comes determined both by its linguistic correctness and by its belonging to the domain of interest.

- A Microplanner, in charge of filling the missing details regarding the concrete implementation document structure, i.e. *how* the information has to be communicated: distribution, referring expressions, level of detail, voice, etc.

- A Surface Realizer, which converts the abstract specification given by the previous stages into a real text, possibly embedded within some medium. It involves traversing the nodal text specification until the final presentation form.

Our approach is based on this generalization, keeping in mind that the document planning –*what* to communicate– is accomplished by the conceptual stages studied previously. In addition to these generic steps, we demand multilingual capabilities and a situation-oriented planning of content, i.e., we want to communicate dangerous or rare events differently than normal developments.

The preliminary implementation follows on work done by Gerber and Nagel [42]. They use Discourse Representation Theory as an abstract framework to identify systematic connections between meaning and linguistic forms. The system consists of three components, all of which need to be adapted when a new language is incorporated, see Fig. 5.2.

High-level predicates from the reasoning stage are eventually converted into surface text, this is, a sequence of words, punctuation symbols, and mark-up annotations to be presented to the user. In order to design the different tasks in the pipe, a set of lemmata has to be first extracted from linguistic corpora on the purposed domain, for each language tackled.

Three tasks are considered: first, lexicalization generates words from predicates with the help of a lexicon, and assigns them a thematic role according to their intended function. Later, these unsorted pieces of knowledge are parsed through a list of DRS construction and transformation rules, which provide structure by progressively reducing free units into constituents of the global sentence. Finally, once a syntax exists, a final step for morphological parsing is applied to make the sentence grammatically and orthographically correct.

A more detailed scheme for the entire process of generation is shown in Fig. 5.3. The sentence "*He is waiting with another pedestrian*" has been generated step by step from logical predicates, for the English language. The three submodules used for NLG (left) are decomposed into specific tasks (center), each one showing its step contribution. The type of information resulting from each task is noted at the right side.

Representation of the discourse

The implementation of semantics for NLG is based on Discourse Representation Theory [63, 61]. This theory aims to provide an abstract framework to systematically represent linguistic information contained in NL sentences, in predicate logic formalism. Semantic relationships are stated by means of DRS. Here, the inverse process is implemented, consisting of the retrieval of NL text from logic predicates, by defining a set of DRS construction and transformation rules for each language.

DRSs are semantic containers which relate referenced conceptual information to linguistic constructions. A DRS consists of a universe of referents and a set of con-

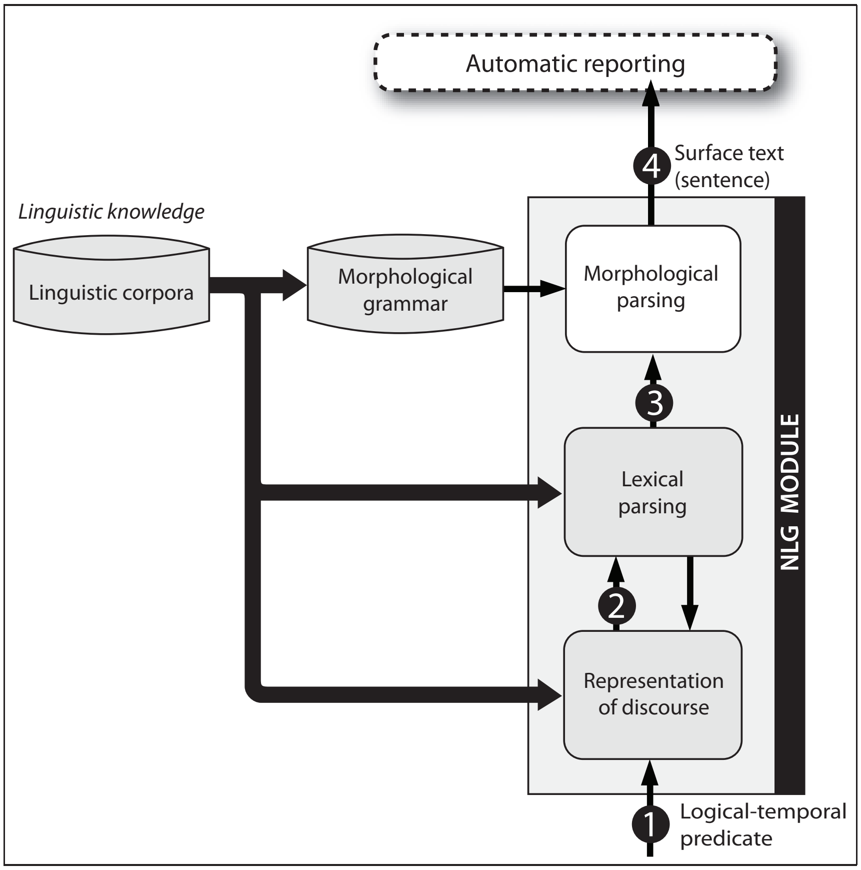


Figure 5.2: Outline of the DRS–NLG module. Darker elements –here, all of them– need modification when adding a new language. Notice that here, REG is accomplished as part of DRS rules.

ditions, which can express characteristics of these referents, relations between them, or even more complex conditions including other DRSs in their definition. These structures contain linguistic data from units that may be larger than single sentences, since one of the ubiquitous characteristics of the DRSs is their semantic cohesiveness for an entire discourse.

When a contextual basis is explicitly provided, the maintenance of the meaning for a discourse, including its cross-references, relations and cohesion can be granted. A particularly interesting and comprehensible example of discourse cohesion is the case of anaphoric pronominalization, which allows the generation of some referring expressions; for instance, we typically discard "The pedestrian waits to cross. The pedestrian crosses", in favor of "The pedestrian waits to cross. S/he crosses". This phenomenon is part of the Referring Expression Generation (REG) problem, i.e., how to refer to an entity depending on the way it has appeared in the discourse up to the moment.

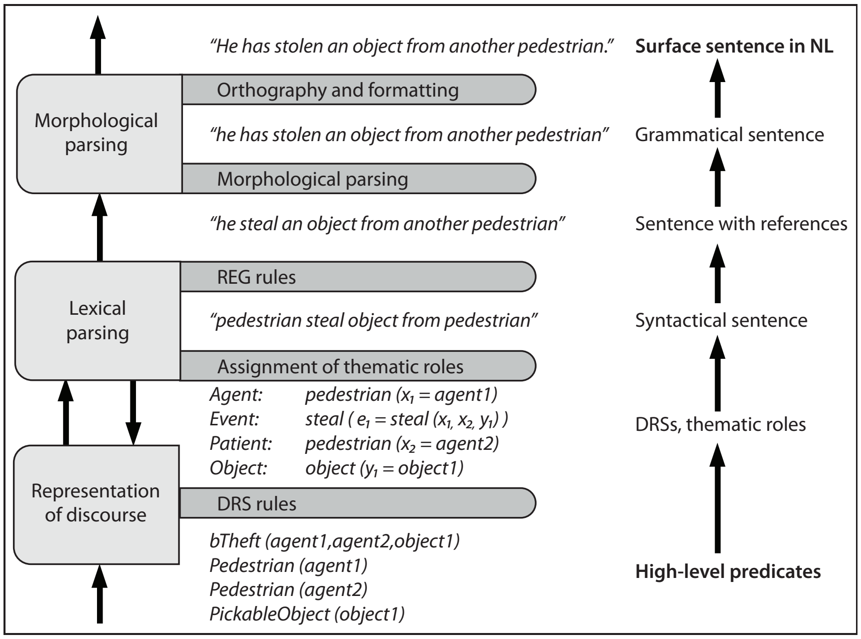

"He has stolen an object from another pedestrian."
Surface sentence in NL
Morphological parsing
Orthography and formatting
"he has stolen an object from another pedestrian"
Grammatical sentence
Morphological parsing
"he steal an object from another pedestrian"
Sentence with references
Lexical parsing
REG rules
"pedestrian steal object from pedestrian"
Syntactical sentence
Assignment of thematic roles
Agent: pedestrian (x_1 = agent1)
Event: steal (e_1 = steal (x_1, x_2, y_1))
Patient: pedestrian (x_2 = agent2)
Object: object (y_1 = object1)
DRSs, thematic roles
Representation of discourse
DRS rules
bTheft (agent1,agent2,object1)
Pedestrian (agent1)
Pedestrian (agent2)
PickableObject (object1)
High-level predicates

Figure 5.3: Step generation of the sentence "He is waiting with another pedestrian" from logical predicates and for the English language.

DRSs point out the cross-references existing among the semantic constituents of a predicate. The classification of linguistically perceived reality into thematic roles (e.g. agent, object, location) is commonly used in contemporary linguistic related applications as a possibility for the representation of semantics, and justifies the use of computational linguistics to describe content extracted by vision processes. In the current implementation, these constituents can be classified as agents, objects, locations, and events/situations. Previously mentioned information about an agent is used to decide upon referenced expressions or full descriptions.

Fig. 5.4 illustrates the way in which a DRS undertakes semantic representation and contextualization. Here, two predicates are validated, which correspond to the observed events *kick vending machine* and *stare at someone*. The first predicate instantiates a DRS, which serves as context for the following asserted facts. Once a new predicate is validated, it instantiates another DRS which merges with that context, thus providing a new context for subsequent facts. The temporal order of the events is stated by including them within time variables ($e_1 \subseteq t_1$), placing these variables in the past ($t_1 \prec n$), and marking precedence ($e_1 \prec e_2$).

DRSs also facilitate the subsequent tasks for sentence generation. The syntactical features of a sentence are provided by the construction rules, which establish the position for the elements of the discourse within a sentence in a particular language. The question of how to address temporal references also arises at the semantic level [62]. There exists certain flexibility for the selection of tenses. This table summarizes a sensible alternative based on the nature of the event:

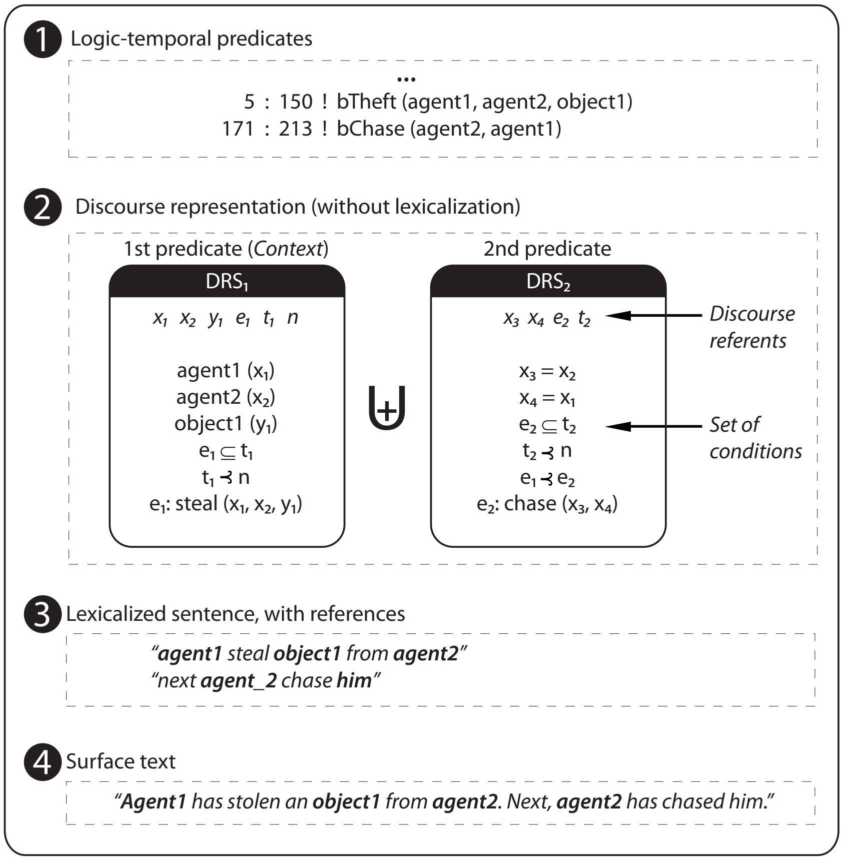

1 Logic-temporal predicates
...
5 : 150 ! bTheft (agent1, agent2, object1)
171 : 213 ! bChase (agent2, agent1)
2 Discourse representation (without lexicalization)
1st predicate (Context)
DRS₁
x₁ x₂ y₁ e₁ t₁ n
agent1 (x₁)
agent2 (x₂)
object1 (y₁)
e₁ ⊆ t₁
t₁ ≺ n
e₁: steal (x₁, x₂, y₁)
⊎
2nd predicate
DRS₂
x₃ x₄ e₂ t₂
x₃ = x₂
x₄ = x₁
e₂ ⊆ t₂
t₂ ≺ n
e₁ ≺ e₂
e₂: chase (x₃, x₄)
Discourse referents
Set of conditions
3 Lexicalized sentence, with references
"agent1 steal object1 from agent2"
"next agent_2 chase him"
4 Surface text
"Agent1 has stolen an object1 from agent2. Next, agent2 has chased him."

Figure 5.4: A pattern DRS allows us to convert a stream of conceptual predicates into a string of textual symbols. The numbers of these step results are linked to Fig. 5.2.

Event type	Tense	Example
Action	Present simple	*stops, turns*
Activity	Present continuous	*is running, is accelerating*
Contextualized event	Present simple	*meets with, leaves*
Behavior interpretation	Uncertain form	*seems to have happened*

A discourse referent n is required for the utterance time of discourse, so that the rest of temporal references can be positioned with respect to it, see Fig. 5.4. Due to the specific goals considered for this system, simple and short sentences are used for effective communication.

```
SPANISH MASCULINE NOUN coche (Object) {
    car(Object)
}
SPANISH REGULAR VERB adelantar (Agent) {
    PREP a (DAT: Object) {
            ATTRIBUTE starting (Object) {
                movefromto(Event,Agent, Object2),
                avoidobstacle(Event2,Agent,Object),
                car(Agent)
            }
    }
}
SPANISH ADVERB ahora (Object) {
    starting(Object)
}
```

Table 5.1
Lexicalization rules provide linguistic form (lemmata) to given predicates or configurations of them.

Lexicalization

Lexicalization is understood as the process of choosing words and syntactic structures to communicate the information in a document plan [106]. In our case, this information is the collection of temporal interpretations in form of logical predicates inferred by the system. Concretely, we have to map a cloud of predicates, now contextualized as DRSs, into words that explain the contents to communicate.

DRS and lexicalization rules are not applied independently, but require of a particular cycle of interaction to accomplish different tasks, like REG. The cycle is performed as follows:

1. DRS construction rules provide an initial structure, by detecting available semantic units.

2. Lexicalization maps the captured semantic units into words.

3. DRS transformation rules affect these words according to their context –e.g, contractions, flexions, REG–.

4. Lexicalization finally substitutes original words by contextualized ones, e.g., pronouns in the case of REG: *A pedestrian meets a pedestrian*→*A pedestrian meets another one*.

This particular cycle becomes particularly difficult to implement for some languages and certain tasks, as it is the case for REG. DRS rules are specific for each language, thus representing a considerable effort in terms of formalization of linguistic phenomena.

$$1.\ \langle\text{“}\textit{go}\text{”}\rangle \begin{bmatrix} \textit{verb} \\ \textit{particip.} \end{bmatrix} \xrightarrow{\text{ENG}} \langle\text{“}\textit{gone}\text{”}\rangle\,[\textit{verb}]$$

$$2.\ \langle\text{“}\textit{meet}\text{”}\rangle \begin{bmatrix} \textit{verb} \\ \textit{particip.} \end{bmatrix} \xrightarrow{\text{ENG}} \langle\text{“}\textit{met}\text{”}\rangle\,[\textit{verb}]$$

$$3.\ \langle\alpha\rangle \begin{bmatrix} \textit{verb} \\ \textit{particip.} \end{bmatrix} \xrightarrow{\text{ENG}} \langle\alpha + \text{“}\textit{ed}\text{”}\rangle\,[\textit{verb}]$$

$$4.\ \langle\text{“}\textit{a}\text{”}\rangle\,[\textit{prep.}] + \langle\text{“}\textit{el}\text{”}\rangle \begin{bmatrix} \textit{determ.} \\ \textit{masc.} \\ \textit{sing.} \end{bmatrix} \xrightarrow[\text{ITA}]{\text{CAT/SPA}} \langle\text{“}\textit{al}\text{”}\rangle \begin{bmatrix} \textit{prep.} \\ \textit{determ.} \\ \textit{masc.} \\ \textit{sing.} \end{bmatrix}$$

$$5.\ \langle\text{“}\textit{de}\text{”}\rangle\,[\textit{prep.}] + \langle\textit{vowel} + \alpha\rangle \xrightarrow[\text{ITA}]{\text{CAT}} \langle\text{“}\textit{d'}\text{”}\rangle\,[\textit{prep.}] + \langle\textit{vowel} + \alpha\rangle$$

$$6.\ \langle\alpha\rangle \begin{bmatrix} \textit{determ.} \\ \textit{sing.} \end{bmatrix} + \langle\textit{vowel} + \beta\rangle \xrightarrow[\text{ITA}]{\text{CAT}} \langle\text{“}\textit{l'}\text{”}\rangle\,[\textit{determ.}] + \langle\textit{vowel} + \beta\rangle$$

Table 5.2
Simple morphological rules in Catalan, English, Italian, and Spanish. More details in the text.

Morphology and surface realization

The surface realization task aims to apply morphological disambiguation at two levels: for each word individually, and for each word considering its neighboring context. The first step applies grammatical attributions like gender or number, stated by the semantic relations previously established by DRSs among the lemmata of discourse. After that, a second set of rules searches for predefined configurations of words that affect the final surface form, due to phenomena like contractions –e.g., *a* + *el* → *al*, in Catalan and Spanish– or order variation. This additional step is indispensable for many languages.

Table 5.2 shows rules included in the grammar for morphological parsing. Rules 1 and 2, in English, reduce the participle tag of a verb for two exceptions, and generate the word form. Rule 3 produces the participle for verbs in a general case. The other rules, for Catalan and Italian, deal with prosodic manipulation: rule 4 covers the contractions of preposition plus determinant, and rules 5 and 6 are for apostrophication, when the following word after certain words starts with a vowel. The syntax of the parser is detailed in Appendix B.

5.3 Ontology-based Natural Language Generation (ONT-NLG)

This section proposes an improvement over the described module for NLG, targeting (i) the easiness of extensibility and flexibility regarding new languages to be implemented, and (ii) the connection of this module to existing ontological resources.

One goal consists of separating technical and linguistic knowledge. This way, native speakers without expert background can add languages by modifying external grammars, without altering the core. In addition, users are allowed to adjust characterizations for each language, metalinguistically. Our motivation for the use of a situated, feature-based approach –*Prototype Theory* from cognitive linguistics– instead of a formal, universal theory –*Discourse Representation Theory*– is that no linguistic rule can be applied universally without having to consider a great amount of exceptions, for one language or another. The following quote expresses this thought:

> *Consider for example the proceedings that we call 'games'. I mean board games, card games, ball games, Olympic games, and so on. What is common to them all? Don't say, "There must be something common, or they would not be called 'games' " - but look and see whether there is anything common to all. For if you look at them you will not see something common to all, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look! Look for example at board games, with their multifarious relationships. Now pass to card games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball games, much that is common is retained, but much is lost. Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear. And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.*
>
> *Philosophical Investigations 66, 1953*
> LUDWIG WITTGENSTEIN (LATER)

In the new scenario we are creating, the properties of a language that are common to another one can then be directly profited, whereas no artificial generalizations will be assumed.

The new layout of the NLG module is presented in 5.5. Linguistic knowledge has been separated from the core processing, so that it can be maintained independently.

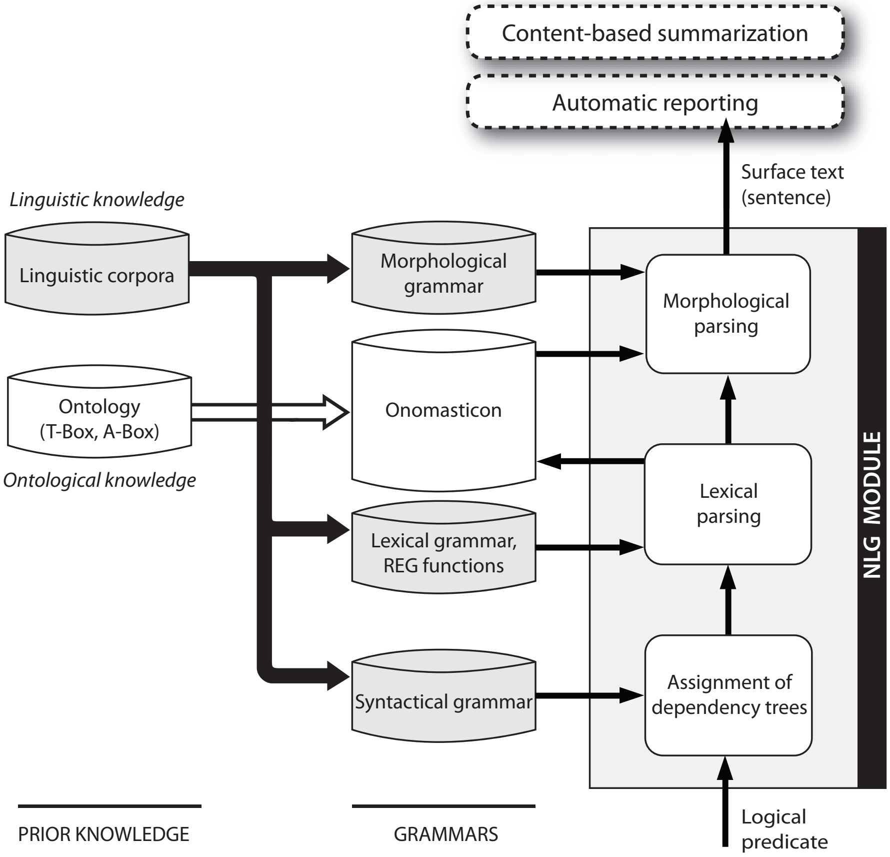


Figure 5.5: The ONT–NLG module enhances previous one. Darker boxes stand for elements that need modifications when incorporating a new language. It is unclear whether REG functions are language-independent; they are for the languages implemented so far.

The process has also been linked to ontological resources, providing a series of integration benefits that will be exploited for additional applications. For instance, this connection couples NLG with the NLU process described in the next section.

Task 1: Assignment of dependency trees

The first task converts an incoming high-level predicate into a tree structure, which gives a unique semantic interpretation to it, and produces a structure for the final surface sentence. Predicate types are linked beforehand to tree templates, whose shapes come predefined by the ontological constraints held by the event; e.g., *is_agent* determines the agent (subject of active sentence) for *wait_with*, see Fig. 5.6. This template-based approach is equivalent to the previous use of DRS template rules.

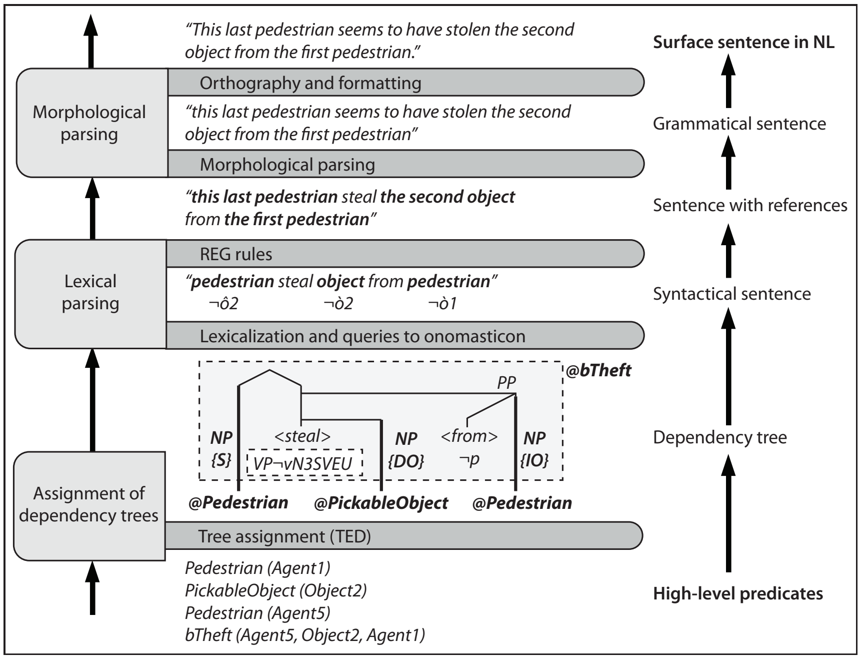

"This last pedestrian seems to have stolen the second object from the first pedestrian."
Surface sentence in NL
Morphological parsing
Orthography and formatting
"this last pedestrian seems to have stolen the second object from the first pedestrian"
Grammatical sentence
Morphological parsing
"this last pedestrian steal the second object from the first pedestrian"
Sentence with references
Lexical parsing
REG rules
"pedestrian steal object from pedestrian"
¬ô2
¬ò2
¬ò1
Syntactical sentence
Lexicalization and queries to onomasticon
@bTheft
PP
NP
{S}
<steal>
VP¬vN3SVEU
NP
{DO}
<from>
¬p
NP
{IO}
Dependency tree
@Pedestrian
@PickableObject
@Pedestrian
Assignment of dependency trees
Tree assignment (TED)
Pedestrian (Agent1)
PickableObject (Object2)
Pedestrian (Agent5)
bTheft (Agent5, Object2, Agent1)
High-level predicates

Figure 5.6: Step generation of a sentence describing a theft, for the English language. This figure extends Fig. 5.5.

Each predicate is linked to an element of the Event T-Box, and its inner fields are as well linked to elements of the Entity and Descriptor T-Boxes. The defined structure reproduces the way in which a situation appears linguistically in the provided corpus. Trees are built by hierarchically using parenthesis to define the conforming nodes. For instance, a parent-child structure is expressed as `(parent(child))`, and a sibling structure would be `(node1)(node2)`. Each node contains a *word* structure –read Appendix B for details– to achieve word aggregation. The inner fields of the predicates will simply be forwarded to the lexicalization process, where an appropriate linguistic structure is assigned to each entity.

An ontological approach offers the possibility to choose *how to communicate the information*, regarding its nature. For instance, it may be desirable to express doubt for uncertain or improbable events, or express continuity for activities that are still on course or under development. A series of predefined linguistic patterns are automatically conferred to the semantic trees by means of tags, depending on the events represented. These rules apply unless a different pattern is chosen for the specific event. Table 5.4 details generally chosen tags for the verbal realizations of the English implementation.

Task 2: Lexicalization and REG

The new lexicalization task also maps semantic elements into linguistic resources that communicate their contents. In this case, units are either words or subtrees, see

```
\\-------------------------|---------------------------------------|
|accelerate(Agent$0)       |(NP{S}:@Agent$0)(VP:<accelerate>¬vN3SR)|;
|appear(Agent$0)           |(NP{S}:@Agent$0)(VP:<appear>¬vN3SR)    |;
|appear(Agent$0,Location$0)|
     (NP{S}:@Agent$0)(VP:<appear>¬vN3SR(PP:<from>(NP:@Location$0)))|;
|back_up(Pedestrian$0)     |
     (NP{S}:@Pedestrian$0)(VP:<back>¬vN3SE(PP:<up>¬p))             |;
\\-------------------------|---------------------------------------|
```

Table 5.3

Assignment of abstract semantic/syntactic tree structures to input high-level predicates. Two predicates sharing the same name but differing in their number of fields generate two different structures.

Type of event	Verbal pattern	Example
Interpretation	vN3SVEU	*This pedestrian seems to be chasing after the third one.*
Contextualization	vN3SR	*S/he has left an object to the ground.*
Activity	vN3SC	*The vehicle is accelerating.*
Action	vN3SR	*The pedestrian walks by the upper sidewalk.*

Table 5.4

Verbal tags ontologically assigned to each event type, for linguistic realization

Fig. 5.7. Whereas the assignment of trees organized elements from the Event T-Box syntactically, lexicalization takes care of ontological elements from the Entity and Descriptor T-Boxes.

The new task involves additional steps. First, *particularizations* must be applied when available, e.g., replacing a general predicate *appear(agent, location)* by *appear(pedestrian, upper_right_side)*, following taxonomical knowledge. Subsequently, lexical realizations are given to a conceptual entities, such as `upper_right` being expanded as “upper right side”, see Fig. 5.7.

The idea of *onomasticon* becomes of great importance regarding REG [37]. REG is known as the task of deciding which expressions should be used to refer to entities, so that the user can easily identify that entity in a given context, see Table 5.5. Traditionally, an onomasticon is a simple repository linking the entities instantiated along the analysis to the possible names one can use to refer to them. In our case,

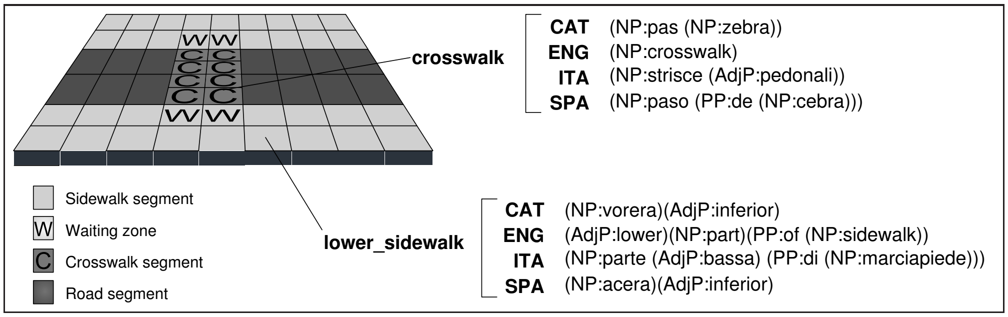
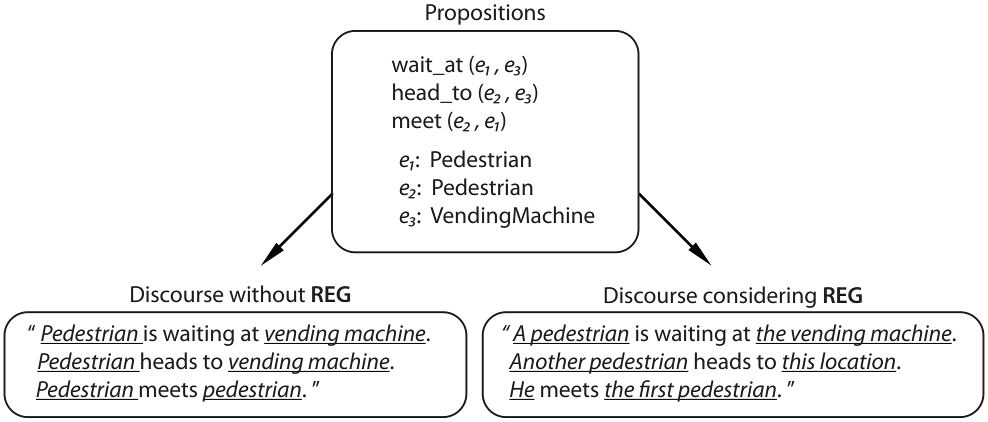

crosswalk
CAT (NP:pas (NP:zebra))
ENG (NP:crosswalk)
ITA (NP:strisce (AdjP:pedonali))
SPA (NP:paso (PP:de (NP:cebra)))
Sidewalk segment
W Waiting zone
C Crosswalk segment
Road segment
lower_sidewalk
CAT (NP:vorera)(AdjP:inferior)
ENG (AdjP:lower)(NP:part)(PP:of (NP:sidewalk))
ITA (NP:parte (AdjP:bassa) (PP:di (NP:marciapiede)))
SPA (NP:acera)(AdjP:inferior)

Figure 5.7: Lexicalization of a priori locations. A linguistic structure is given to each semantic region of the scenario, for each language.



wait_at (e_1, e_3)
head_to (e_2, e_3)
meet (e_2, e_1)

e_1: Pedestrian
e_2: Pedestrian
e_3: VendingMachine

" *Pedestrian* is waiting at *vending machine*. *Pedestrian* heads to *vending machine*. *Pedestrian* meets *pedestrian*. "

" *A pedestrian* is waiting at *the vending machine*. *Another pedestrian* heads to *this location*. *He* meets *the first pedestrian*. "

Table 5.5

The REG task avoids possible ambiguities when identifying entities in a generated discourse.

an onomasticon is extended by tracking instances along the discourse, allowing the system to answer questions like: *has it ever been instantiated?*, *more than once?*, *are there other instances of the same concept?*, *was it the central entity in the last sentence generated?*, or *was the last instance definite*?

The proper combination of these REG cases allows the ONT–NLG module to choose the most appropriate referring expression, like *an [entity]*, *a new [entity]*, *the [entity]*, *this last [entity]*, *the second [entity]*. For example, if we have seen a car in the scene previously, and a new agent of type car appears, we use "*a new car*"; otherwise, if none of the vehicles or other agents seen was specifically a car, we use simply "a car", thus highlighting the class instead of the actual instantiation.

REG situations have been abstracted with independence of the language, to account the different linguistic references useful for our application. It is possible that

REG needs being revisited when a new language is implemented, since they may have distinct lexicalization needs from the ones implemented at the moment. So far, though, the generic REG models have covered needs of reference for the current languages implemented. Next, a formal definition of this problem and the proposed solution is presented; Appendix B contains extended technical details for the parsing REG rules implemented.

We define the *instantiation operator* as follows:

Let

$\mathcal{C}$	be a given context, and
$\mathcal{E} = \{y\}_M$	be a set of entities.
$\mathcal{A}^{\mathcal{C}} = \{(i\ x_n^{\mathcal{C}})y\}_{M \times N_m}$	be the set of instances of $\mathcal{E}$ in $\mathcal{C}$,

The n-th instantiation of the m-th entity y_m in context $\mathcal{C}$ is formalized as $(i\ x_n^{\mathcal{C}})y_m$, where $x_n^{\mathcal{C}} \in \mathcal{A}^{\mathcal{C}}, y_m \in \mathcal{E},\ \forall n = 1, 2 \dots N,\ \forall m = 1, 2 \dots M$.

Now we will study the REG casuistry.
Let

$\mathcal{A}_m^{\mathcal{C}} = \{(i\ x_n^{\mathcal{C}})y_m\}_{N_m}$	be the subset of instances of y_m in $\mathcal{C}$, and
$\mathcal{A}^{\mathcal{C}}\|_{\mathcal{P}_k}$	be the subset of $\mathcal{A}^{\mathcal{C}}$ which have been instantiated at $\mathcal{P}_k$, where $\mathcal{P}_k$ is the k-th generated proposition and $k = 1, 2 \dots K$.

Finally, let **A,B,. . . ,F** be test functions over the instance $(i\ x_n^{\mathcal{C}})y_m$, defined as:

A :	$n > 1$	The instance is a subsequent reference [106].
B :	$n == N_m$	The instance has been the last instantiated one.
C :	$(i\ x_{N_m}^{\mathcal{C}})y_m \rightarrow \texttt{Def}$	Last appearance of the entity was definite.[2]
D :	$\mid \mathcal{A}^{\mathcal{C}}\|_{\mathcal{P}_K} \mid > 1$	Last proposition contained more than one instance.
E :	$\mid \mathcal{A}_m^{\mathcal{C}}\|_{\mathcal{P}_K} \mid = 1$	Last proposition contains only one instance of y_m.
F :	$\mid \mathcal{A}^{\mathcal{C}}\|_{\mathcal{P}_J} \mid > 1$, $\mid \mathcal{A}_m^{\mathcal{C}} \mid = 1$	There were several instances of y_m, but only one left.
G :	N_m	Number of apparitions of the instance.

Depending of the different values resulting from the application of the test functions over the instance $(i\ x_n^{\mathcal{C}})y_m$, a set of possible REG situations have been defined, see Table 5.6. Each and every one of these situations is associated to a certain REG tag, which has to be defined in the categories file. At the moment, these rules solve the linguistic needs which were tackled for this implementation of the NL text generator. Nevertheless, this is not a finished solution for the problem: the necessary test functions should be better analyzed, in order to be consistent with the universality of the REG situations for each language, and the list of these situations should be extended as required.

[2] →`Def` is the result of having classified the instance in question as definite.

REG Feature	Tag	**A**	**B**	**C**	**D**	**E**	**F**	**G**	Example
`REG.Undefinite`	ó	×						1	*a*
`REG.NewUndefinite`	õ	×						>1	*a new*
`REG.Definite`	ò		✓	×	×				*the*
		✓	×					1	*the*
`REG.AppearedLast`	ô		✓	✓	×				*s/he*
`REG.AppearedLastMultiple`	ò2		✓		✓	✓			*this last*
`REG.Nth`	ò4		✓		✓	×		n	*the n-th*
		✓	×					1	*the n-th*
`REG.Remaining`	ö						✓		*the remaining*
`REG.AlreadyReferred`	o	—	—	—	—	—	—	—	

Table 5.6
Table of REG features according to the information provided by the onomasticon

Task 3: Morphology and surface realization

Finally, the morphological and surface realization process involves mapping the specification of a text into a surface text form, i.e. a sequence of words, punctuation symbols, and mark-up annotations to be presented to the end-user [106]. In practice, it consists of applying parsing techniques to modify either independent words (verb inflections or conjugations, plurals) or words depending of their surrounding context (contractions, vowel adjacency, prosodic effects). In the example of Fig. 5.9, the third person of the verb has been conjugated; similarly, this step also updates tenses ("*leave*"→"*has left*") and changes words in context ("*a agent*"→"*an agent*"). As a result of the morphological process, a rich semantic/syntactic tree structure with referred expressions and morphological forms is generated. The linearization of the tree nodes and a final addition of orthographical and formatting information provides a final surface form for the end-user.

5.4 Natural Language Understanding (ONT-NLU)

The NLU module has to choose the most appropriate interpretation out of a set of possible ones, given a textual input in NL. In our case, the ontology specifies the domain of validity undertaken by the universe of possible user queries, and makes it possible to reduce them to a handleable space of situations. In addition, to avoid excessive ambiguity when resolving the meaning of the inputs, this module accepts uniquely single (not compound) sentences from the end-users.

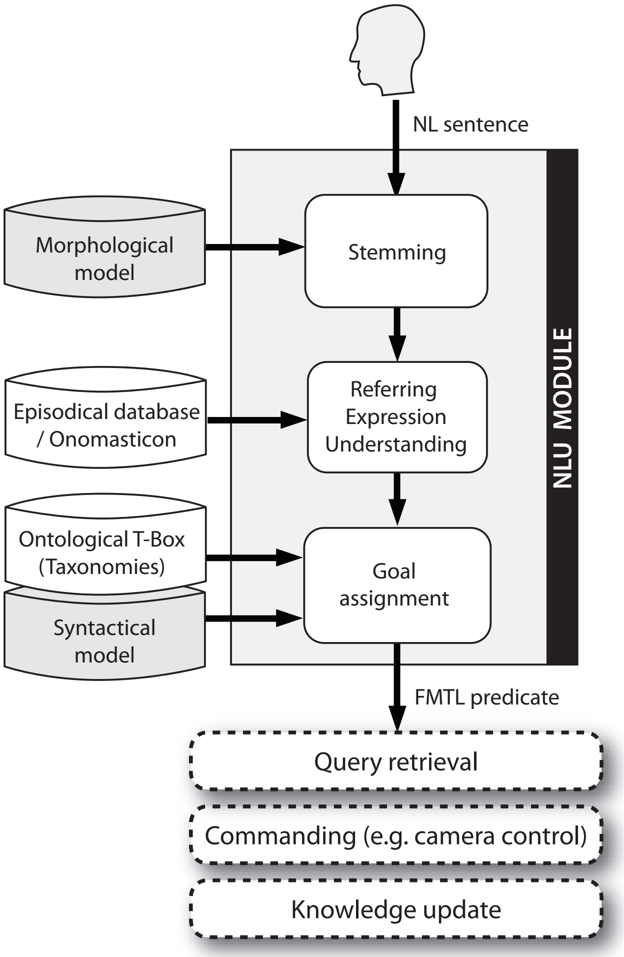

NL sentence
Morphological model
Stemming
Episodical database / Onomasticon
Referring Expression Understanding
NLU MODULE
Ontological T-Box (Taxonomies)
Syntactical model
Goal assignment
FMTL predicate
Query retrieval
Commanding (e.g. camera control)
Knowledge update

Figure 5.8: Scheme of the NLU module. Sentences written by the user are individually converted into conceptual predicates.

The general operations conducted by the NLU module are shown in Fig. 5.8. First, each sentence of the user is processed by a stemming algorithm, and its contents are linked to concepts from the global ontology. After that, the specific context of the sentence is found by relating all referring expressions to their corresponding entity instances, with the help of the so-called onomasticon. Finally, the interpreted sentence is analyzed at a syntactic/semantic level, and its contents are assigned to the most suitable action predicate in order to generate virtual agents in the scene. These three steps are explained in detail next.

Stemming

The first step of the NLU module consists of mapping the surface form of a natural language sentence into a more simplified and structured form, by removing those elements that are not significant for a semantic evaluation and maintaining only word stems in canonical forms, i.e. as lemmata. This process is accomplished by means of a traditional rule-based parsing technique.

The transformation of the sentence into a structure of annotated lemmata is again done at two levels: individually and by context. Individual word tagging extracts linguistic characteristics for a word and annotates them as a chain of grammatical tags, used to disambiguate the sentence. In addition, stop-words (e.g., determinants) are removed from the sentence. At a contextual level, parsing is carried out by additionally considering the neighborhood of a particular word; this case is fundamental to detect collocations or expressions referring to a unique ontological concept, e.g., *"vending machine"*. Some examples of tagging rules are shown next.

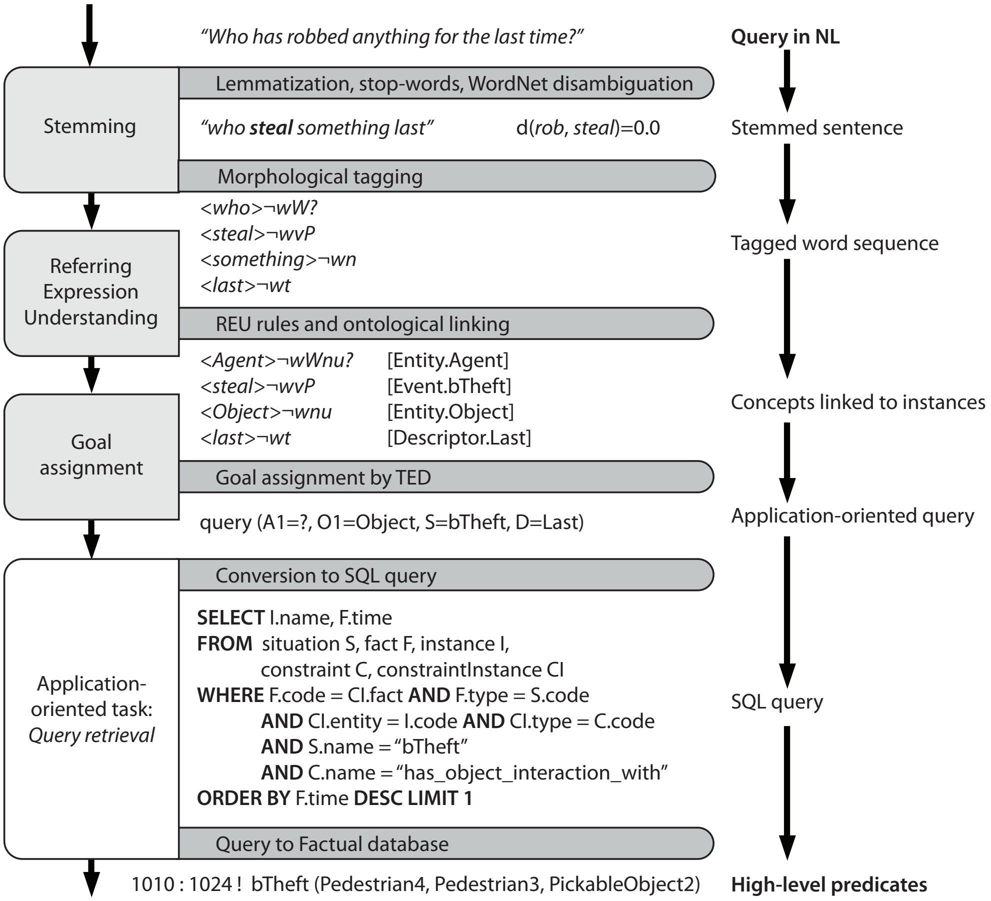
```
|  [<com_>] <up>          |  Ĵwv            @Appear         |;
|  <him>                  |  {P}ĴwrN3S      @Entity         |;
|  <towards>              |  Ĵwp                            |;
|  <lower> <left> <side>  |  <lowleft>Ĵwn   @LowerLeftSide  |;
```

In the examples, "¬" denotes the start of a chain of grammatical tags, and "@" denotes a link to the ontology. The first line detects the two words of any non-past form of *"come up"* and links them to the default expression in the ontology, i.e., *Appear*, also marking it as a word (w) and verb (v). The second example tags the 3rd person singular pronoun (rN3S) *"him"* as an ontological entity, also classifying it as a patient (P). The third example tags a preposition as such (p). Finally, the fourth line merges the expression of a predefined location of the scenario into a single noun (n).

A basic goal of this process is to link each lemma to a concept from the ontology, so that the possible interpretations of the input sentence are reduced to those admissible by the defined models. While tackling this problem, an NLU module has to deal with unknown terms or expressions, for which no conceptual knowledge is made explicit within the system. In order to augment the recognition rate of words in the domain, and to additionally avoid scaling the linguistic models to cover them all, the reliance on very generic databases (e.g., common-knowledge or linguistic repositories) opens possibilities of learning or adaptation. In our case, further lexical disambiguation is accomplished relying on the WordNet lexical database [34].

WordNet is a linguistic database that groups the words of a language into sets of concepts called *synsets*, which manifest the semantic proximity of these words. Such synsets are in turn related to each other by parentive relationships of hypernymy and hyponymy[3], and contain other valuable information such as use cases and definitions of meaning. Currently, English WordNet includes approximately 155.000 words and

[3] An *hyponym* is a term that presents all the semantic characteristics of a more general term –an *hypernym* to the first term–.


"Who has robbed anything for the last time?"
Query in NL
Stemming
Lemmatization, stop-words, WordNet disambiguation
"who **steal** something last"
d(*rob*, *steal*)=0.0
Stemmed sentence
Morphological tagging
<who>¬wW?
<steal>¬wvP
<something>¬wn
<last>¬wt
Tagged word sequence
Referring Expression Understanding
REU rules and ontological linking
<Agent>¬wWnu? [Entity.Agent]
<steal>¬wvP [Event.bTheft]
<Object>¬wnu [Entity.Object]
<last>¬wt [Descriptor.Last]
Concepts linked to instances
Goal assignment
Goal assignment by TED
query (A1=?, O1=Object, S=bTheft, D=Last)
Application-oriented query
Application-oriented task: *Query retrieval*
Conversion to SQL query
SQL query
Query to Factual database
1010 : 1024 ! bTheft (Pedestrian4, Pedestrian3, PickableObject2)
High-level predicates

```
SELECT I.name, F.time
FROM  situation S, fact F, instance I,
      constraint C, constraintInstance CI
WHERE F.code = CI.fact AND F.type = S.code
      AND CI.entity = I.code AND CI.type = C.code
      AND S.name = "bTheft"
      AND C.name = "has_object_interaction_with"
ORDER BY F.time DESC LIMIT 1
```

Figure 5.9: Step results for the NLU process in a case of query retrieval. The concepts linked to words are either *Facts* or *Entity Instances* from the factual database, as seen in Fig. 4.11.

117.000 synsets, structured into 4 lexical categories –nouns, verbs, adjectives, and adverbs–.

In order to measure the semantic distance of an unknown word to terms known by the system, the unknown word is compared to the list of taxonomical concepts that share the same lexical category. A distance value is retrieved using semantic metrics based on relationships such as synonymy and hypernymy. New candidates are evaluated to determine the ontological nature of an unknown word; as a result, the word is linked to a number of domain concepts that can explain it.

Referring Expression Understanding

An *onomasticon* is a repository that keeps track of the different linguistic expressions that correspond to the same entity in a discourse. In our case, an extended onomasti-

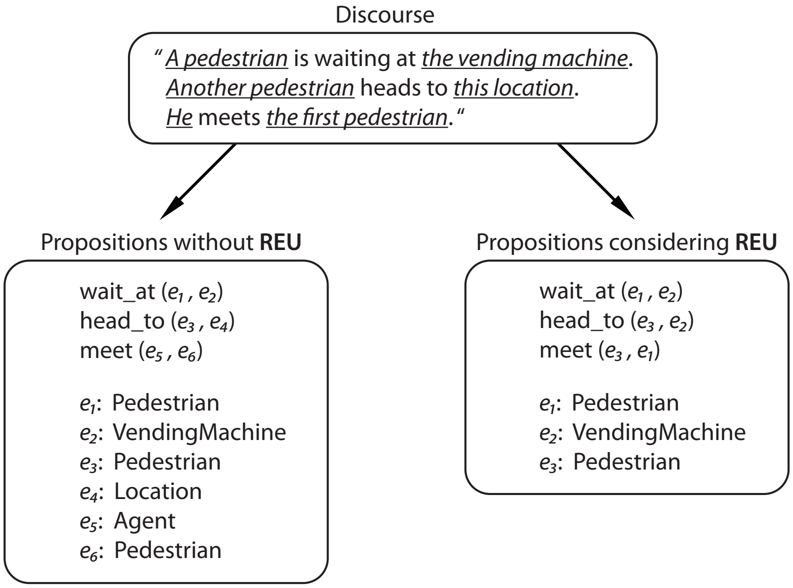


Table 5.7

The REU task keeps track of entities in a discourse, linking propositions to their implied entity instances. Omitting this task leads to semantic ambiguity.

con is additionally aware of how the instantiation of entities has been done: whether a certain entity has ever been instantiated, whether this has been done in the last sentence, or how many instances of each entity does the discourse have at any moment, for example. This module is of great importance to accomplish the REU, a task to decide *which* entities in the discourse are referred *by which* textual expressions; we aim to identify them according to previous information available. Table 5.7 shows the importance of the REU task in the understanding of natural language.

The identification of entities by referring expressions is carried out by managing a set of test functions over the existing instances, which evaluate cases like: (1) the instance has been referred at least once, (2) it has been the last instantiation, (3) the last appearance of the same entity was definite, (4) the last proposition contained more than one instance, (5) the last proposition contains only one instance of the same entity, or (6) the instance has appeared more than once during the discourse. Depending on the answers to these questions, end-users refer to one or another entity using expressions like *"a person"* (¬1,¬6), *"a new person"* (¬1,6), *"the person"* (1,¬2,¬6 or 2,¬3,¬4), *"s/he"* (2, 3,¬4), or *"this last person"* (2,4,5).

Assignment of adjacency trees

Following the idea of hypothesis management, the NLU module links textual sentences to their most accurate interpretations in the domain, in form of predicates related to scene concepts and instances. Once a proper formatting has been applied, an input sentence is analyzed through a sequence of 3 processes [37]: first, a morphological parser tags words with linguistic features depending on the context of apparition, and a syntactic/semantic parser builds a dependency tree out of the tagged sentence. Secondly, the resulting tree with ontological references is assigned to the most related query predicate from a collection of patterns. Finally, the obtained predicate is used to query the factual database of indexed occurrences. The process is detailed next.

The semantic part of the analysis already starts with the word tagging process: the lexical models attach domain concepts to words that potentially refer to them. Hence, there are two issues to solve, since (i) a word can be linked to several concepts, e.g., word "turn *left*" (concept *OrientationDescriptor*) and "*left* entrance" (*Location*); and (ii) each concept may also have many words attached to it, as for the words "person", "pedestrian", or "walker" and the concept *Pedestrian*. Parsing rules solve the first ambiguity. Regarding the second issue, a robust system must be able to understand not modeled words, i.e., to sensibly link unknown words to a domain concepts. To this end, we rely on the WordNet lexical database [34] to retrieve lists of closely related words, using semantic metrics based on synonymy and hypernymy. New word candidates are evaluated to determine the nature of the unknown word. As a result, the word is linked to a number of concepts that can explain it.

Next, a dependency tree is built with the help of syntactical rules, which first identify the heads of phrase classes and then recursively nest words and phrases hierarchically. The resulting tree is then compared to a collection of tree patterns by computing a semantically-extended Tree Edit Distance (TED) [18], see Fig. 5.10. In order to compute the TED, the concepts at the leaves of the pattern trees are aligned to those from the test tree, and the TED evaluates the coincidence of each concept: it penalizes strongly the absences, penalizes the generalizations proportionally to the number of levels to the test concept, and does not penalize at all when the test concept matches or particularizes the pattern one. For example, the concept *Car* augments the distance with pattern tree 2 having *Pedestrian* at the corresponding leaf, but specializes the general concept *Vehicle* in the same position of pattern 3 with distance zero.

The pattern tree with lowest distance to the test tree is decided as the most valid interpretation, and the fields of its associated predicate are particularized with specific information from the sentence. These predicates, called *goal predicates*, have been restricted to the 4 different types shown next, towards a practical implementation. The main elements to retrieve from NL sentences are especially situations (S), and also agents (A), objects (O), locations (L), and time expressions (t), which can be refined by ontological descriptors.

Query type	NL example	Equivalent goal predicate
Assert	*Has anybody run after a robbery?*	`Assert{A=Agent,S=Run,t=After(Theft)}`
Count	*How many robberies have happened?*	`Count{S=Theft}`
Query	*When has an agent run by the road?*	`Query{A=Pedestrian,S=Run,L=Road}`
List	*Which vehicles have been observed?*	`List{A=Vehicle,t=Before(Theft)}`

A final step adapts each goal predicate pattern to the relational language used for the factual database, in this case SQL. The retrieval process returns the entries that

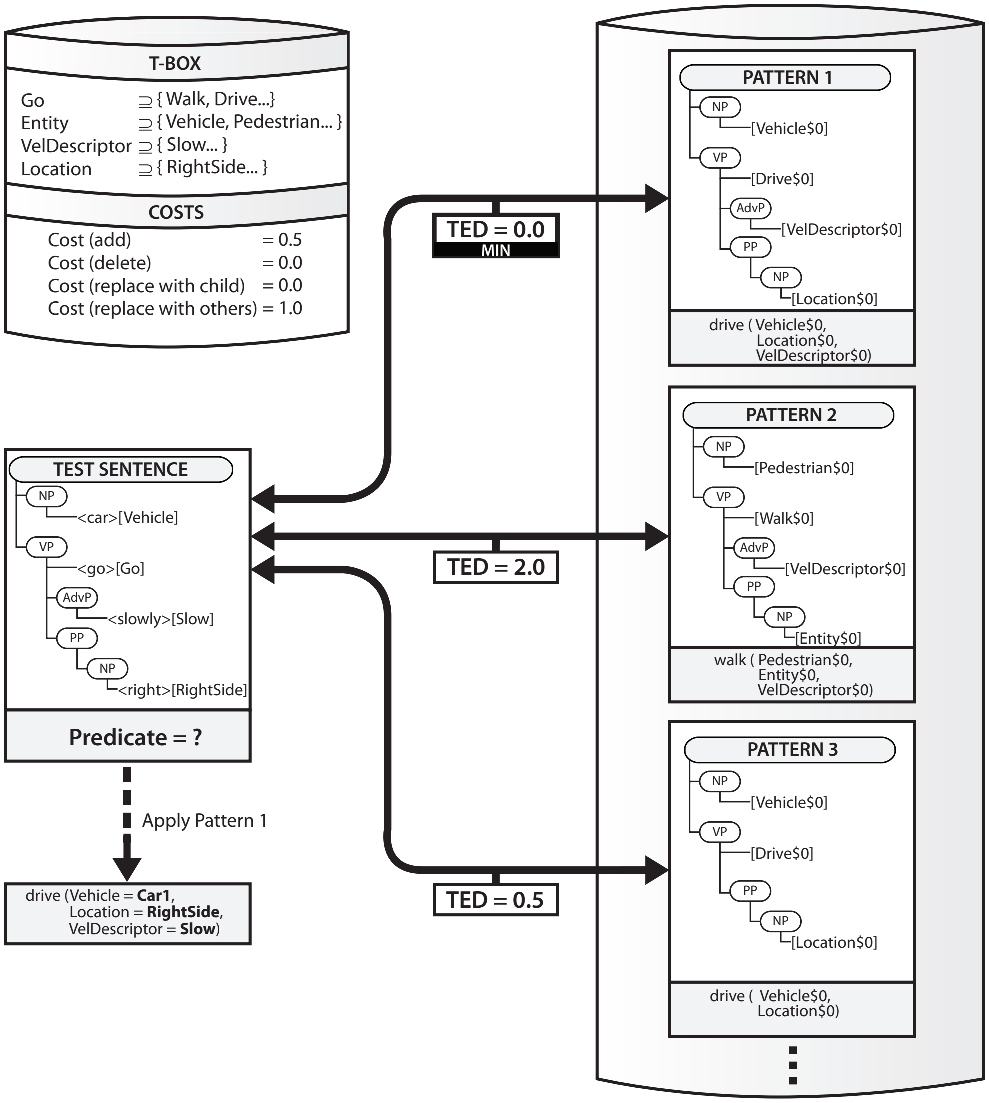


Figure 5.10: The test sentence is compared to a collection of pattern trees, each one associated to an abstract predicate. The predicate of that pattern with a lowest TED is specialized with information from the sentence.

satisfy the NL query of the end-user, along with the interval of the video sequence corresponding to the event index. Some examples of NL-based retrieval are presented in the next section, along with the rest of the experimental results.

5.5 Synthetic generation and augmentation of scenes (SA)

It is desirable for modeling formalisms to not only represent and recognize model instances, but also facilitate their synthetic generation. This section demonstrates that the presented framework can be adapted to synthesize image sequences with behavioral content. Our field of work suggests three potential applications of interest:

- Generating synthetic image sequences that represent temporal occurrences expressed by logical predicates.
- Augmenting real scenes with virtual agents, whose behavior is linguistically defined by end-users.
- Synthesizing complex environments to evaluate aspects like the tracking system (e.g., crowded scenes) or the behavioral models (e.g., detecting inconsistencies).

The first task, *synthetic scene generation*, enables the system to recreate virtual scenes representing the detected events and behaviors. In addition, it becomes a very visual and unequivocal way to evaluate the understanding of the scene by the system, compared to our own. Virtual scenes that are equivalent to real ones –in terms of contents– implicate an immense compression of the information, which is reduced to a list of temporally-valid predicates. The second problem, *scene augmentation*, is solved by combining virtual scenes with real recordings. In our case, we aim to generate virtual agents that accomplish goals and react to real occurrences of the scene, in order to have sophisticated means of simulating and evaluating modeled behaviors of our framework. Both tasks can be applied to performance evaluation.

Fig 5.11 shows the outline of the Scene Augmentation (SA) module, which makes it possible for the system to generate synthetic behavioral agents that react to real developments. The SA module also produces fully synthetic sequences by omitting real data and using conceptual models for the scenario, thus accomplishing the first (simpler) objective. In addition, a proper use of the NLU module allows end-users to describe, using NL, the behaviors of agents for scene synthesis or augmentation.

The main characteristics of the suggested approach for scene augmentation are: (i) end-users describe the behavior of virtual agents in form of NL textual commands, and (ii) virtual agents react in real-time to real occurrences within the scene.

In order to accomplish these objectives, several steps are progressively followed. In the first place, real world occurrences must be interpreted, for which we use the SGT-based interpretation framework described in the previous chapter. On the other hand, virtual agents are given a series of goals in form of NL descriptions of events, and converted into predicates by means of the NLU module presented in Section 5.4. Hence, high-level predicates for *interpretations* and *goals* are inputs to the SA module, whose components are described next.

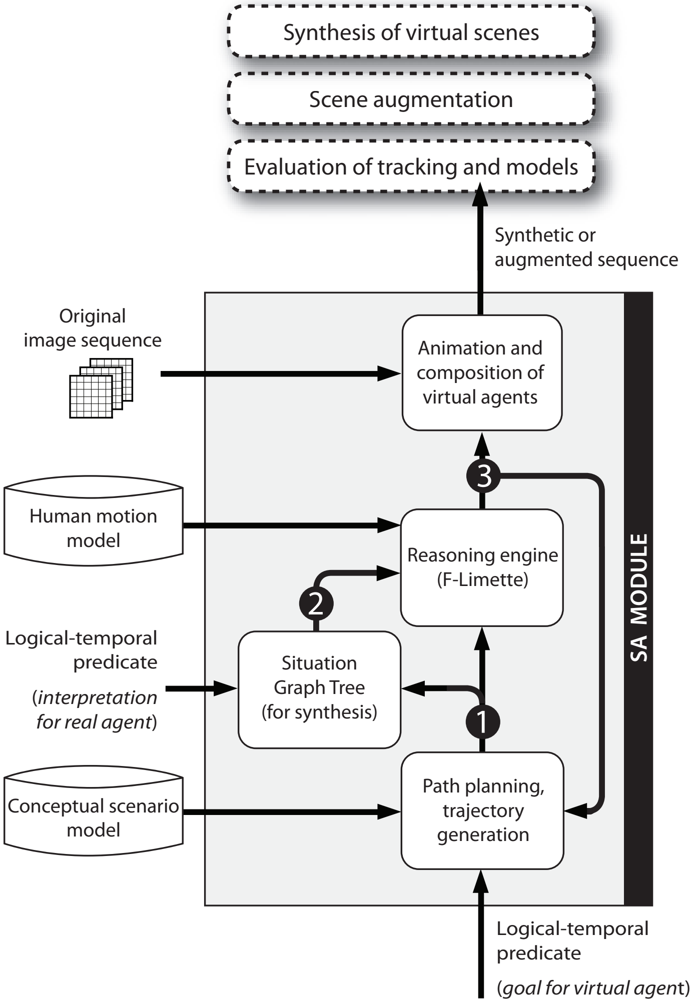

Synthesis of virtual scenes
Scene augmentation
Evaluation of tracking and models
Synthetic or augmented sequence
Original image sequence
Animation and composition of virtual agents
3
Human motion model
Reasoning engine (F-Limette)
2
Logical-temporal predicate
(interpretation for real agent)
Situation Graph Tree (for synthesis)
1
Conceptual scenario model
Path planning, trajectory generation
SA MODULE
Logical-temporal predicate
(goal for virtual agent)

Figure 5.11: The SA module augments real sequences with virtual elements, linguistically described by end-users. Fig. 5.13 gives examples of predicates found at the three highlighted positions of the diagram.

Path planning and trajectory generation

End-users write textual commands for event generation that are interpreted by the NLU module, producing goals for the virtual agents in form of high-level predicates. These goals are of the form `cross_street(Agent1)`, `leave(Agent1, left)`, `follow(Agent2, Agent1)`, `talk(Agent2, Agent1)`, thus referring to existing elements in the scene. The primary step for a virtual agent is to move towards the destination entailed by its given goal, e.g., the opposite sidewalk, the left side of the scene, or the location of *Agent2*. However, while moving to accomplish its goal, this agent may have to react to other agents, virtual or real, and also to the semantic properties of the environment.

The family of predicates `go_to` compute the minimum path –in terms of a sequence of contiguous scenario segments– to go from the agent's current position to the location implied by the goal predicate. If no further restrictions are imposed, the shortest and straightest path is taken. Nevertheless, semantic restrictions come associated by the behavioral models, e.g., pedestrians should cross roads only by pedestrian crossings, and if there are no vehicles coming. Or for example, pedestrians will take the less crowded of two alternative paths.

The SGT formalism also allows us to model and apply the behavioral restrictions for virtual agents. This time, goal predicates are decomposed into a series of partial objectives involving connected paths. Consider the following decomposition:

```
cross_street   →   go_to_closest_waiting_line
                   reduce_speed
                   go_to_the_other_waiting_line
                   leave_scene
```

Each partial objective is formalized as a situation scheme in Fig. 5.12 (a), where a crowd of agents appearing randomly from a sidewalk side are told to cross to the other side. This implementation of the behavior makes them approach to the closest crosswalk, reduce speed, reach to the other side and leave by the closest sidewalk exit.

Basic intermediate objectives are changes of location –`go_to(Agent, Position)`–, changes of action –`change_to_performing(Agent,Action)`–, and changes of velocity. A position can be fixed or related to another agent, and the current agent's action defines its instantaneous posture, as seen farther ahead. Intermediate objectives are accomplished by a time-step generation of trajectories, in form of instantaneous `has_status` predicates:

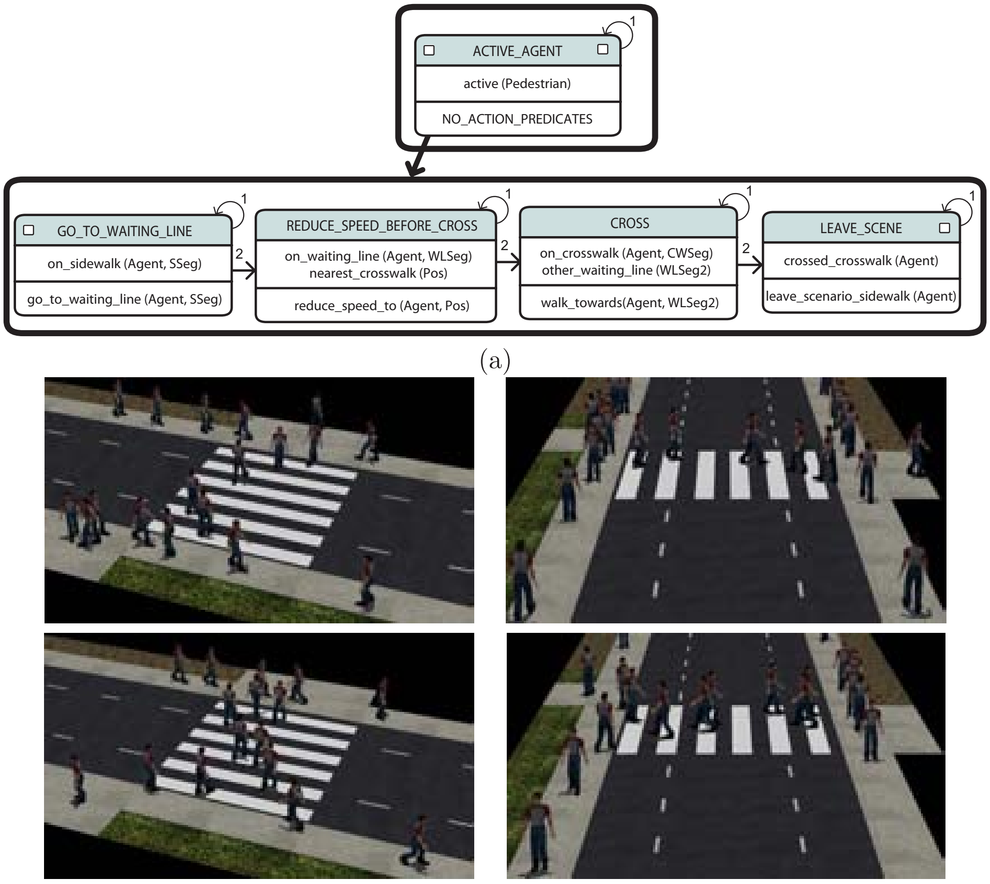
$$t \;\; ! \;\; has\_status\ (agent,\ x_t,\ y_t,\ \theta_t,\ v_t,\ action_t,\ p_t) \tag{5.1}$$

where (x_t, y_t) are computed from the previous location (x_{t-1}, y_{t-1}) using generic motion models; *action* can be either *stand*, *walk*, or *run*, depending on the velocity; and $p \in [0, 1]$ is a frame-incremented parameter that cyclically covers the possible postures within an action, as commented in subsequent sections.

Reactive behaviors

SGTs are able to decompose goals into intermediate objectives, but they also provide virtual agents with capabilities to react against external stimuli, making them autonomous in restricted environments. The SA module recomputes trajectories derived from intermediate objectives every frame, to determine the best way to achieve them. The reason for this is that certain situations –e.g., a sudden obstacle on the way– force agents to adapt to the environment and follow alternative plans.

To this end, a priori trajectories are also adapted by SGTs. In the example of Fig. 5.13, the current position of the virtual agent is analyzed by an SGT to determine possible obstacles on the way, and in that case affects the original trajectory for the next step. As a result, a detour in the trajectory is generated in real-time.


ACTIVE_AGENT
active (Pedestrian)
NO_ACTION_PREDICATES
1
GO_TO_WAITING_LINE
on_sidewalk (Agent, SSeg)
go_to_waiting_line (Agent, SSeg)
2
REDUCE_SPEED_BEFORE_CROSS
on_waiting_line (Agent, WLSeg)
nearest_crosswalk (Pos)
reduce_speed_to (Agent, Pos)
CROSS
on_crosswalk (Agent, CWSeg)
other_waiting_line (WLSeg2)
walk_towards(Agent, WLSeg2)
LEAVE_SCENE
crossed_crosswalk (Agent)
leave_scenario_sidewalk (Agent)

(a)

(b)

Figure 5.12: (a) SGT for the *cross_street* behavior, used to test interactions with the environment. (b) Simulation of a crowd of virtual agents performing the *cross_street* behavior.

Animation and composition

Once the spatio-temporal status of the agent has been determined, it has to be rendered in a virtual or augmented scene. In the first place, the rendered appearance of an agent varies according to the step of its action, see Fig. 5.14. Each animated action has an *aSpace* learned [44], whose parameter p determines the posture to be shown at every frame. The *aSpaces* used for scene augmentation are mainly *stand*, *walk*, and *run*.

The final step is the composition of an augmented image sequence, containing both real and virtual agents processed at time t. In order to give consistence and realism, the occlusions among scene elements need to be handled. For each pair of agents (A_1, A_2) in the scene having positions r_1 and r_2, respectively, we compute their distances to the position of the camera r_C as $d_1 = \overline{r_1 r_C}$, $d_2 = \overline{r_2 r_C}$. Agents

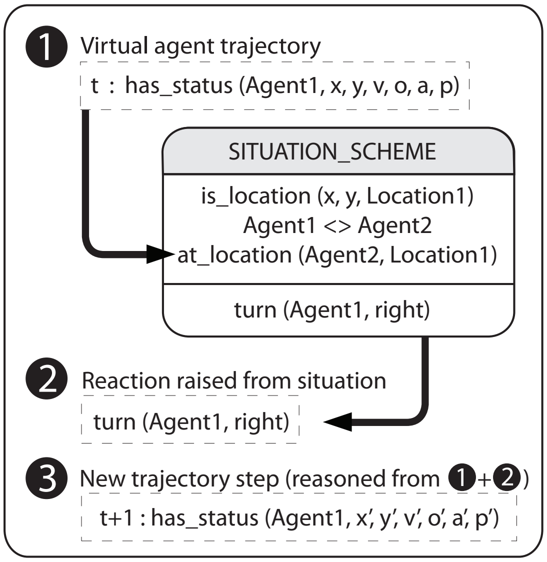

1 Virtual agent trajectory
t : has_status (Agent1, x, y, v, o, a, p)
SITUATION_SCHEME
is_location (x, y, Location1)
Agent1 <> Agent2
at_location (Agent2, Location1)
turn (Agent1, right)
2 Reaction raised from situation
turn (Agent1, right)
3 New trajectory step (reasoned from 1+2)
t+1 : has_status (Agent1, x', y', v', o', a', p')

Figure 5.13: SGTs modify dynamically the trajectory of virtual agents, according to the behavioral models defined. These predicates are found at the different positions of the SA module, as depicted in Fig. 5.13.

are sequentially superimposed over the background, sorted by their distance to the camera –larger first–.

Fig. 5.15 shows a real and a virtual agent having distances d_1 and d_2 to the camera, respectively. Since $d_1 < d_2$, the real agent occludes the other one. In case of pure scene synthesis, only the relative positions among virtual agents are considered, and a virtual model of the scenario is used as background.

5.6 Applications of ONT–NLG

Two main applications have been developed using the ONT–NLG module: the automatic description of video events in multiple languages, and the automatic summarization of such reports based on selectable content. Ontologies play an important role in both of them, regarding language extensibility and adaptability, and content management and centralization.

Automatic multilingual reporting

The incorporation of new languages into the ONT–NLG comes facilitated by the systematic ontological design that we have described. We have identified the problems of the original DRS–NLG, inspired by the Angus2 system, problems that happened especially at the REG and morphological levels. The ONT–NLG presented implements Catalan, Basque, English, Italian, Spanish, and Turkish languages so far, i.e., languages from the Indo-European family (Germanic and Romance), from the Turkic-

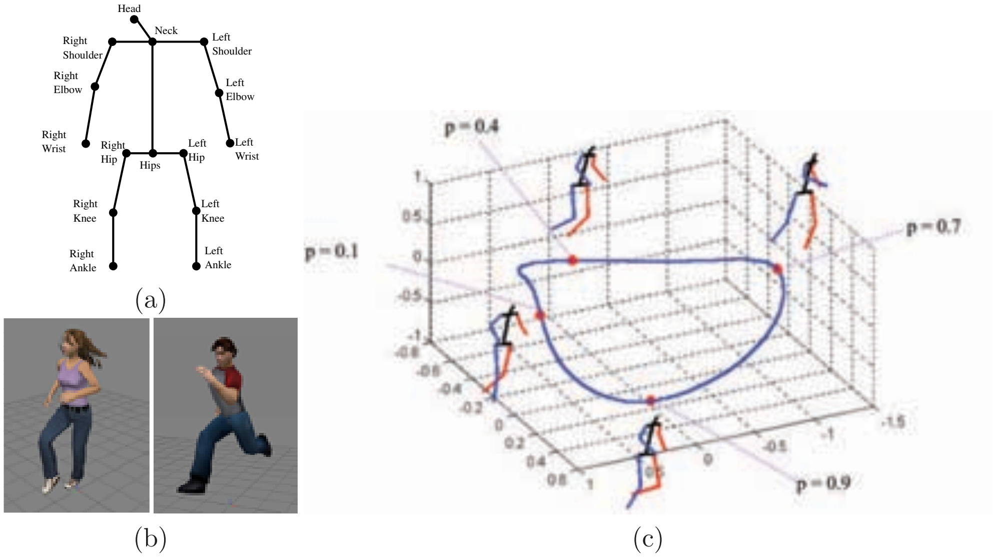


Figure 5.14: (a) Generic human model (stick figure). (b) Rendered models performing *dance* and *run* actions. (c) Animation in the *aRun aSpace* [44]: variations of p evolve a posture temporally, modifying the disposition of body-parts.

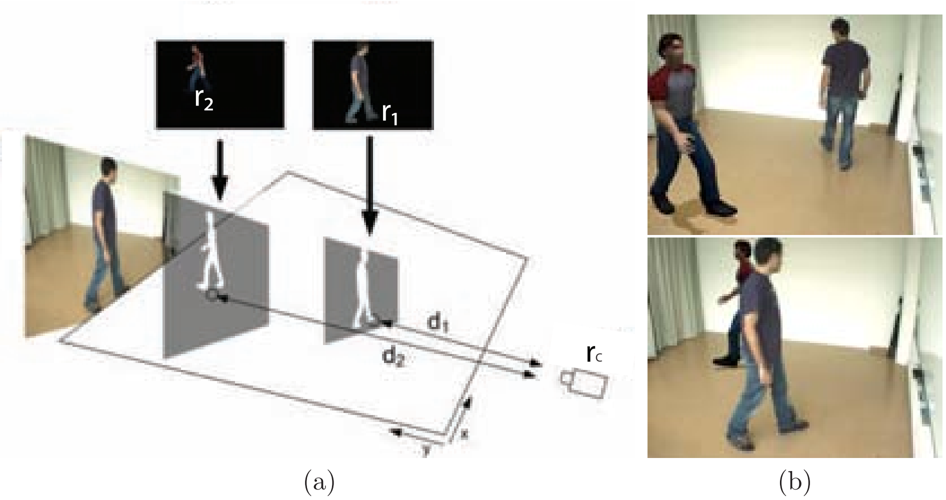


Figure 5.15: (a) To compose an augmented scene, the agents are added sequentially according to their distance to the camera. (b) Result of the composition.

Altaic families (Turkish), and even language isolates like Basque, which is the last remaining pre-Indo-European language in Western Europe, not linkable to any of its neighbors in the continent[4]. The previous implementation of DRS–NLG was addi-

[4] Larry Trask, *The History of Basque*. Routledge, 1997. ISBN 0-415-13116-2.

tionally tested on Czech, German, and Japanese, resulting in a significative range of languages. Although it cannot be assured that NLG can be assumed in any possible language with our current system, this does demonstrate the consistency of the system for multilingual generation.

This section presents automatic reports for 3 different video surveillance sequences: *ZEBRA*, *HERMES-Outdoor*, and *HERMES-Indoor*, in Tables 5.8, 5.9, and 5.10, respectively. For a better understanding, the English results have been included in every case. The average measured time for generating a single sentence has been 3 milliseconds, using a Pentium D at 3.20GHz with 2GB RAM.

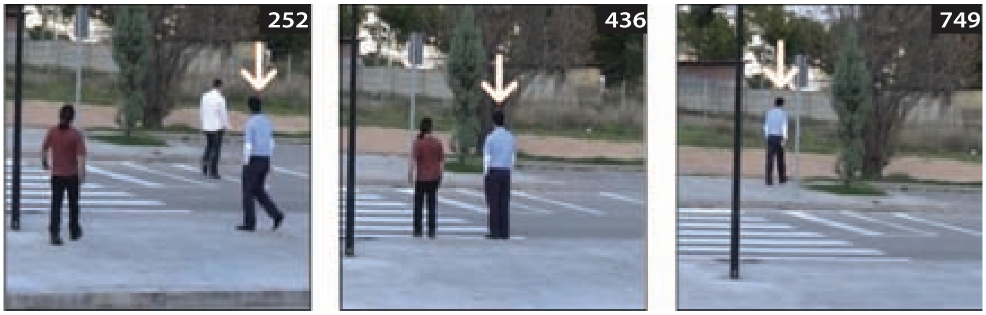
Languages like Basque naturally take into account the presence of contextual objects in the sentence to construct it; for instance, in Basque we find a difference between *“A person has left by an exit”*→*“Pertsona hau sarreratik atera da”* and *“A person has forgotten an object”*→*“Pertsona honek objektua ahaztu du”*, where *hau* and *honek* are used according to the nature of the contextual linguistic units. Similar phenomena can be found for some of the other languages. The use of mechanisms in natural languages to explicitly distinguish the type of event–entity relations reinforces the validity of our choice of terminological organization of knowledge, in cognitive terms.

The synthetic results presented have been compared to the corpus produced by the 30 English native speakers already described in Chapter 4, see Table 5.16. Less than one third of the subjects are members of a computer science department, and none of them has NL processing background. Subjects were told to describe both sequences in a natural and linguistically correct manner, using the expressions they considered most suitable. The quantitative evaluation carried out in the previous chapter compared statistically the synthetic annotations to the most common user descriptions. However, a qualitative evaluation is required to examine the naturality and expressivity of the results.

A qualitative evaluation allows us to detect differences between the set of facts detected by the subjects and the one generated by the system. On the other hand, we also want to learn the mechanisms of reference used, and which kind of words, expressions, and connectors are being most employed. These have been compared to our choices. When considering the list of facts to compare to the inputs, facts having closely related meanings have been gathered together, e.g., *“notice”*–*“realize”*, or ‘ ‘*run after”*–*“chase”*–*“chase after”*.

- A requirement for *economy* is deduced: when one states *“A man walks down the sidewalk”*, there is no need to include *“A man appears”*. Also, there is no need to state that a person is *“bending”* when picking up an object; it is obvious when the object is known to be on the ground.

- The greater difference regarding expressiveness happens when the subjects deduce the intentions of the agents by the context, using common sense. For instance, *“He waves his hands in amazement that the car didn’t stop”* or *“He seemed somewhat hesitant”*. Sometimes, the following situations in the scene are anticipated, like *“A person is walking to the zebra crossing to meet someone”*. These constructions are very useful to conduct the discourse.

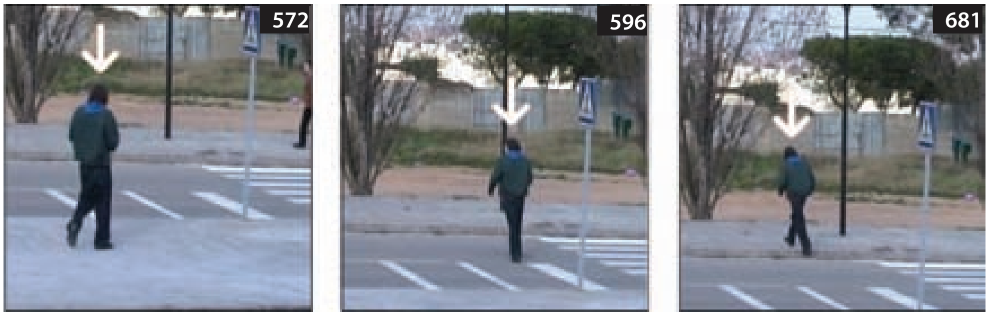


CATALAN

203 ! Lo vianant surt per la part inferior dreta.
252 ! Va per la vorera inferior.
401 ! S'espera per creuar.
436 ! S'està esperant amb un altre vianant.
506 ! Creua pel pas zebra.
616 ! Va per la vorera superior.
749 ! Se'n va per la part superior dreta.

ENGLISH

203 ! The pedestrian shows up from the lower right side.
252 ! S/he walks on the lower sidewalk.
401 ! S/he waits to cross.
436 ! S/he is waiting with another pedestrian.
506 ! S/he enters the crosswalk.
616 ! S/he walks on the upper sidewalk.
749 ! S/he leaves by the upper right side.



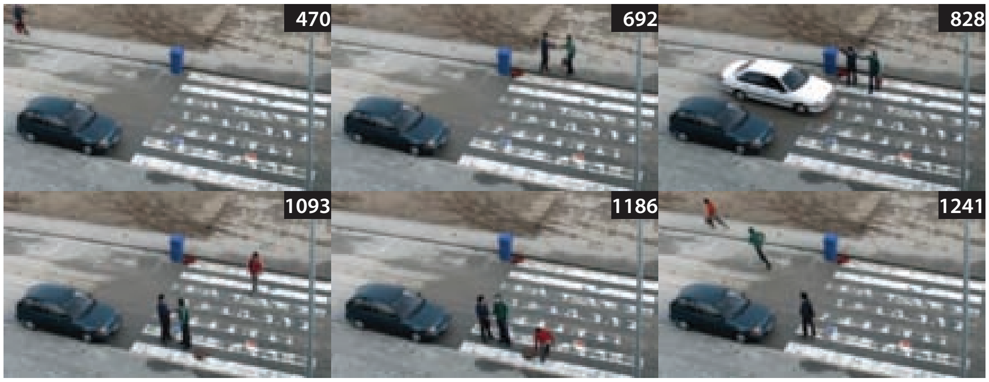
ENGLISH

523 : The pedestrian shows up from the lower left side.
572 : S/he walks on the lower sidewalk.
596 : S/he crosses the road carelessly.
681 : S/he walks on the upper sidewalk.
711 : S/he leaves by the upper left side.

SPANISH

523 : El peatón aparece por la parte inferior izquierda.
572 : Camina por la acera inferior.
596 : Cruza sin cuidado por la calzada.
681 : Camina por la acera superior.
711 : Se va por la parte superior izquierda.

Table 5.8
Catalan, Spanish, and English NL reports generated for some of the pedestrians appearing in *ZEBRA*.



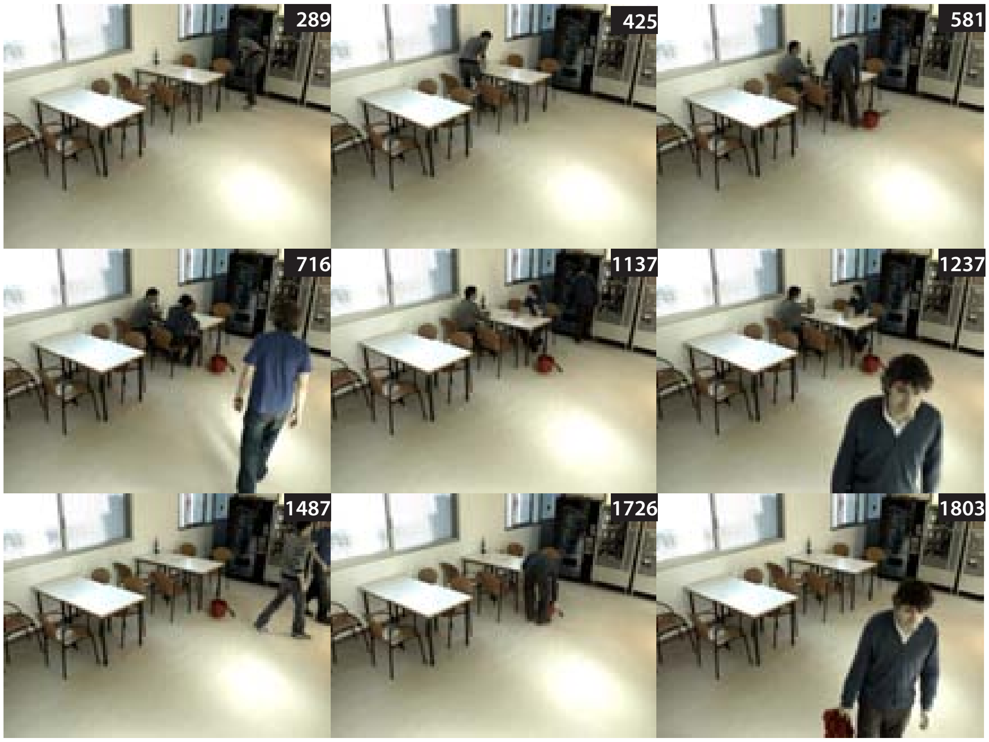
ENGLISH

470 ! A pedestrian appears by the upper left side.
492 ! The pedestrian is walking by the upper sidewalk.
583 ! S/he has turned right in the upper part of the crosswalk.
591 ! S/he has stopped in the same place.
615 ! S/he has left an object.
630 ! A new pedestrian appears by the upper right side.
642 ! The pedestrian is walking by the upper sidewalk.
656 ! The first pedestrian is walking by the same place.
687 ! The object seems to have been abandoned in the upper part of the crosswalk.
692 ! The first pedestrian has met the second one there.
799 ! The second pedestrian enters the crosswalk.
806 ! A vehicle appears by the left.
810 ! The first pedestrian enters the crosswalk.
822 ! It seems that a danger of runover between this pedestrian and the vehicle occurred.
825 ! This last pedestrian has stopped.
828 ! The vehicle is braking up.
828 ! It seems that a danger of runover between the second pedestrian and the vehicle occurred.
838 ! This last pedestrian has backed up.
842 ! S/he has stopped.
852 ! The vehicle is accelerating.
862 ! A new vehicle appears by the left.
872 ! The first vehicle has exited by the right.
891 ! The remaining vehicle gives way to the first pedestrian.
891 ! This last vehicle gives way to the second pedestrian.
896 ! This last pedestrian enters the crosswalk.
906 ! The first pedestrian enters the same place.
939 ! The vehicle is accelerating.
1000 ! The second pedestrian has stopped in the lower part of the crosswalk.
1006 ! The first pedestrian has stopped in the same place.
1018 ! The vehicle has exited by the right.
1033 ! A new pedestrian appears by the upper right side.
1049 ! The pedestrian is walking by the upper sidewalk.
1054 ! The second pedestrian has left a new object in the crosswalk.
1078 ! The third pedestrian has turned left in the upper part of the crosswalk.
1093 ! S/he enters the crosswalk.
1168 ! S/he has turned right.
1186 ! S/he picks up the second object.
1211 ! This last pedestrian is running by the road.
1220 ! It seems that this pedestrian has stolen the second object to the first pedestrian.
1241 ! The second pedestrian seems to be chasing after the third one.
1276 ! The third pedestrian has exited by the upper left side.

TURKISH

470 ! Bir yaya sol üst tarafta belirir.
492 ! Yaya kaldırımın yukarı tarafında yürür.
583 ! Yaya geçidinin üst tarafından sağa döner.
591 ! Orada durur.
615 ! Yere bir cisim bırakır.
630 ! Başka bir yaya sağ üst tarafta belirir.
642 ! Yaya kaldırımın yukarı tarafında yürür.
656 ! Birinci yaya orada yürür.
687 ! Yaya geçidinin üst tarafında cisim yere bırakılır.
692 ! Bu son yaya orada ikinci yaya ile buluşur.
799 ! İkinci yaya yaya geçidine girer.
806 ! Bir araç sol tarafta belirir.
810 ! Birinci yaya yaya geçidine girer.
822 ! Bu yaya ve araç arasında bir ezilme tehlikesi yaşanır.
825 ! Bu son yaya durur.
828 ! Araç bozulur.
828 ! İkinci yaya ve bu araç arasında bir ezilme tehlikesi yaşanır.
838 ! Bu son yaya geri çekilir.
842 ! Durur.
852 ! Araç ivmelenir.
862 ! Başka bir araç sol tarafta belirir.
872 ! Birinci araç sağ taraftan terk eder.
891 ! Araç birinci yayaya yol verir.
891 ! Bu son araç ikinci yayaya yol verir.
896 ! Bu son yaya yaya geçidine girer.
906 ! Birinci yaya aynı yere girer.
939 ! Araç ivmelenir.
1000 ! İkinci yaya yaya geçidinin üst tarafında durur.
1006 ! Birinci yaya orada durur.
1018 ! Araç sağ taraftan terk eder.
1033 ! Başka bir yaya sağ üst tarafta belirir.
1049 ! Yaya kaldırımın yukarı tarafında yürür.
1054 ! İkinci yaya yaya geçidine başka bir cisim bırakır.
1078 ! Üçüncü yaya yaya geçidinin üst tarafından sola döner.
1093 ! Yaya geçidine girer.
1168 ! Sağa döner.
1186 ! İkinci cismi yerden alır.
1211 ! Bu son yaya yolda koşar.
1220 ! Bu yaya birinci yayadan ikinci cismi çalar.
1241 ! İkinci yaya üçüncü yayayı kovalar.
1276 ! Üçüncü yaya sol üst taraftan terk eder.

Table 5.9
English and Turkish NL reports generated for *HERMES-Outdoor*.



BASQUE

200 ! Pertsona lehen sarreratik agertu da.
270 ! Makina saltzailearen aurrean gelditu da.
289 ! Erabili egin du.
425 ! Bigarren mahaian eseri da.
501 ! Pertsona berria bigarren sarreratik agertu da.
581 ! Bigarren mahaian eseri da.
602 ! Objektua utzi du.
716 ! Pertsona berria lehen sarreratik agertu da.
882 ! Bigarren mahaian eseri da.
1073 ! Bigarren pertsona zutitu egin da.
1137 ! Makina saltzailea ostikoz jo duela dirudi.
1237 ! Lehen sarreratik atera da.
1237 ! Pertsona honek kafetegian lehen objektua ahaztu duela dirudi.
1395 ! Lehen pertsona zutitu egin da.
1424 ! Hirugarren pertsona zutitu egin da.
1487 ! Bigarren sarreratik atera da.
1501 ! Lehen pertsona leku beretik atera da.
1501 ! Pertsona honek bigarren mahaian objektu berria ahaztu duela dirudi.
1655 ! Pertsona berria lehen sarreratik agertu da.
1726 ! Lehen objektua jaso du.
1803 ! Azken pertsona hau lehen sarreratik atera da.

ENGLISH

200 ! A person appears by the first entrance.
270 ! S/he stops in front of the vending machine.
289 ! S/he uses it.
425 ! S/he sits down at the second table.
501 ! A new person appears by the second entrance.
581 ! S/he sits down at the second table.
602 ! S/he leaves an object.
716 ! A new person appears by the first entrance.
882 ! S/he sits down at the second table.
1073 ! The second person stands up.
1137 ! S/he seems to kick the vending machine.
1237 ! S/he leaves by the first entrance.
1237 ! It seems that this person has abandoned an object.
1395 ! The first person stands up.
1424 ! The third person stands up, too.
1487 ! S/he leaves by the second entrance.
1501 ! The first person leaves by the same place, too.
1501 ! It seems that this person has abandoned a new object.
1655 ! A new person appears by the first entrance.
1726 ! S/he picks up the first object.
1803 ! S/he leaves by the first entrance.

Table 5.10

Basque and English NL reports generated for *HERMES-Indoor*.

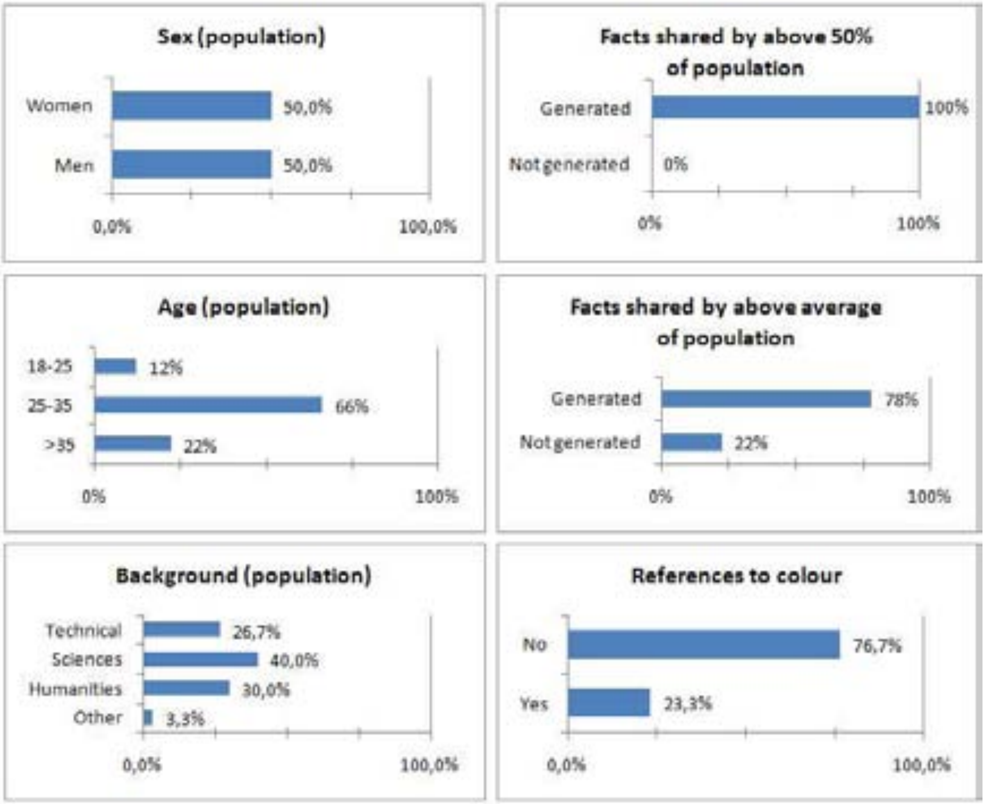


Figure 5.16: Statistics about the NL generation experiment, for English and the outdoor sequence. The population consisted of 30 subjects from different backgrounds. The left column contains information about the population, the right column shows quantitative results about the evaluation and comparison with the facts used.

- One of the main tasks lacking in the described generation system is the aggregation of simple sentences into more complex and expressive ones, using mechanisms such as coordination or subordination. This has the main advantage of emphasizing certain elements of the discourse. For instance, *"After crossing the street they stop, one of them puts his bag on the ground, and while they are talking, another guy comes and snatches the bag"* prioritizes the object left over the crossing and the theft over the talk.

- The use of certain adverbs, adjectives, and other complementary words has been seen as helpful towards a richer discourse: *"nearly hit by an oncoming vehicle"*, *"jumps back in surprise"*, *"move back slightly"*, *"they only crossed the street half-way"*, among others.

In addition, it is also interesting to notice that just about one quarter of the population has included color references to support their descriptions. Most of these (above 70%

of them) use a single reference, for the “white car”, which is the only agent with a very distinctive color.

Content-based summarization

Content-based summarization is a direct application of an ontologically founded NLG. Since each concept to be described has been instantiated by the ontology, the generation module can easily filter certain conceptual information (events, entities, locations, etc.) to be converted into linguistical terms. As a result, the final text shows only the specified content, avoiding unrelated information.

Table 5.11 contains 4 texts in English describing the `HERMES-Outdoor` scene. The long text on the left is a maximally detailed generation, incorporating all *Status*, *Contextualization*, and *Interpretation* events. On the right side we find three summaries: in the upper one, *Status* have been discarded. The middle one additionally discards *Contextualization*, and the lower one includes only sentences implying a specific object. Similarly, reports can be restricted to concrete agents or locations.

5.7 Applications of ONT-NLU

In the case of the ONT-NLU module, ontological resources become a fundamental channel to conduct the input of end-users to actions or responses that can be managed by the system. As the capabilities of the system increase, a centralized repository of structured knowledge facilitates this task. Two applications are shown for this module: the retrieval of video contents by means of NL queries, and a small adaptation to perform visual storytelling, with the help of the SA module.

NL query retrieval

This section evaluates the capability of the system to retrieve video content from NL queries. The objective is to correctly map any potentially valid linguistic utterance into the limited domain of the application, and also decide when this is not possible, making the query invalid for that domain.

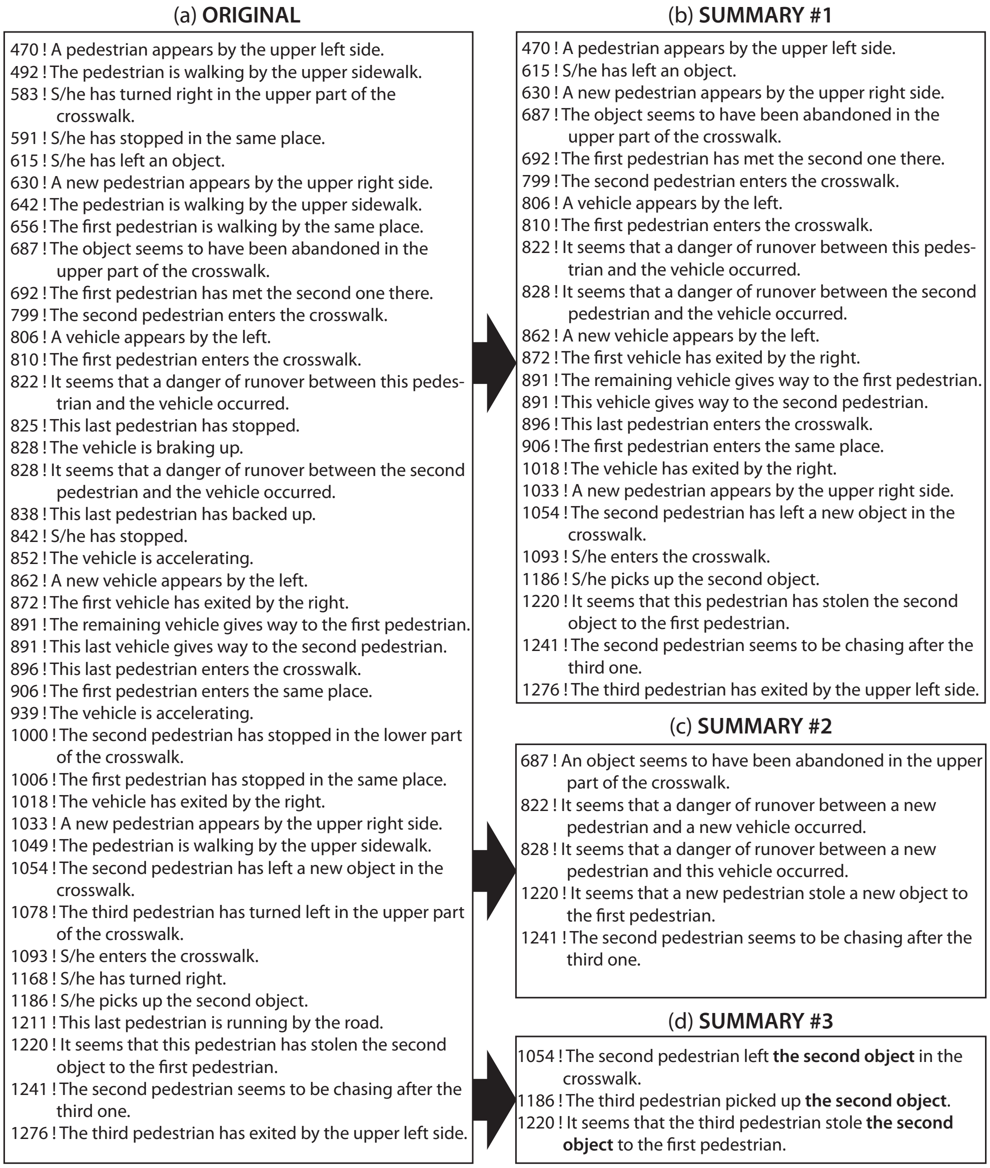
One of the critical issues of this application is the handling of unknown words or expressions by the system. Several existing Java APIs have been considered to facilitate the exploitation of the WordNet repository, viz `JAWS`[5], `JWNL`[6], and `RiTa`[7]. This last one has been chosen for the intuitive and resourceful list of functions that presents, in addition to its more accurate results when computing semantic distances. Algorithm 5.7 details how the NLU module determines the closest concept to an input word, and how new rules are incorporated to the system. In practice, the system proposes the new rules to the user before incorporating them.

The behavior of the system in front of unknown terms is shown in Table 5.13. A Pentium 4 at 2.4 GHz and 1 GB RAM has been used to accomplish these tests. The

[5] http://lyle.smu.edu/~tspell/jaws/index.html

[6] http://sourceforge.net/projects/jwordnet/

[7] http://www.rednoise.org/rita/wordnet/



(a) **ORIGINAL**

470 ! A pedestrian appears by the upper left side.
492 ! The pedestrian is walking by the upper sidewalk.
583 ! S/he has turned right in the upper part of the crosswalk.
591 ! S/he has stopped in the same place.
615 ! S/he has left an object.
630 ! A new pedestrian appears by the upper right side.
642 ! The pedestrian is walking by the upper sidewalk.
656 ! The first pedestrian is walking by the same place.
687 ! The object seems to have been abandoned in the upper part of the crosswalk.
692 ! The first pedestrian has met the second one there.
799 ! The second pedestrian enters the crosswalk.
806 ! A vehicle appears by the left.
810 ! The first pedestrian enters the crosswalk.
822 ! It seems that a danger of runover between this pedestrian and the vehicle occurred.
825 ! This last pedestrian has stopped.
828 ! The vehicle is braking up.
828 ! It seems that a danger of runover between the second pedestrian and the vehicle occurred.
838 ! This last pedestrian has backed up.
842 ! S/he has stopped.
852 ! The vehicle is accelerating.
862 ! A new vehicle appears by the left.
872 ! The first vehicle has exited by the right.
891 ! The remaining vehicle gives way to the first pedestrian.
891 ! This last vehicle gives way to the second pedestrian.
896 ! This last pedestrian enters the crosswalk.
906 ! The first pedestrian enters the same place.
939 ! The vehicle is accelerating.
1000 ! The second pedestrian has stopped in the lower part of the crosswalk.
1006 ! The first pedestrian has stopped in the same place.
1018 ! The vehicle has exited by the right.
1033 ! A new pedestrian appears by the upper right side.
1049 ! The pedestrian is walking by the upper sidewalk.
1054 ! The second pedestrian has left a new object in the crosswalk.
1078 ! The third pedestrian has turned left in the upper part of the crosswalk.
1093 ! S/he enters the crosswalk.
1168 ! S/he has turned right.
1186 ! S/he picks up the second object.
1211 ! This last pedestrian is running by the road.
1220 ! It seems that this pedestrian has stolen the second object to the first pedestrian.
1241 ! The second pedestrian seems to be chasing after the third one.
1276 ! The third pedestrian has exited by the upper left side.

(b) **SUMMARY #1**

470 ! A pedestrian appears by the upper left side.
615 ! S/he has left an object.
630 ! A new pedestrian appears by the upper right side.
687 ! The object seems to have been abandoned in the upper part of the crosswalk.
692 ! The first pedestrian has met the second one there.
799 ! The second pedestrian enters the crosswalk.
806 ! A vehicle appears by the left.
810 ! The first pedestrian enters the crosswalk.
822 ! It seems that a danger of runover between this pedestrian and the vehicle occurred.
828 ! It seems that a danger of runover between the second pedestrian and the vehicle occurred.
862 ! A new vehicle appears by the left.
872 ! The first vehicle has exited by the right.
891 ! The remaining vehicle gives way to the first pedestrian.
891 ! This vehicle gives way to the second pedestrian.
896 ! This last pedestrian enters the crosswalk.
906 ! The first pedestrian enters the same place.
1018 ! The vehicle has exited by the right.
1033 ! A new pedestrian appears by the upper right side.
1054 ! The second pedestrian has left a new object in the crosswalk.
1093 ! S/he enters the crosswalk.
1186 ! S/he picks up the second object.
1220 ! It seems that this pedestrian has stolen the second object to the first pedestrian.
1241 ! The second pedestrian seems to be chasing after the third one.
1276 ! The third pedestrian has exited by the upper left side.

(c) **SUMMARY #2**

687 ! An object seems to have been abandoned in the upper part of the crosswalk.
822 ! It seems that a danger of runover between a new pedestrian and a new vehicle occurred.
828 ! It seems that a danger of runover between a new pedestrian and this vehicle occurred.
1220 ! It seems that a new pedestrian stole a new object to the first pedestrian.
1241 ! The second pedestrian seems to be chasing after the third one.

(d) **SUMMARY #3**

1054 ! The second pedestrian left **the second object** in the crosswalk.
1186 ! The third pedestrian picked up **the second object**.
1220 ! It seems that the third pedestrian stole **the second object** to the first pedestrian.

Table 5.11

Report of the HERMES-Outdoor scene in English, (a) considering no summarization, (b) discarding very basic events, (c) showing only domain interpretations, and (d) informing about a particular scene element.

```
Require: $th \in (0, 1)$
Ensure: $d \leftarrow min\ (semantic\_distance\ (w, c))$ and $d \in [0, 1]$
  if not $exists\_lemmatization\_rule\ (w)$ then
    $Candidates \leftarrow \emptyset$
    for $c \subset \mathcal{T}$ do
      if $type(w) = type(c)$ then
        $d = semantic\_distance\ (w, c)$
        if $d \leq th$ then
          $Candidates \leftarrow Candidates\ \cup\ < c, d >$
        end if
      end if
    end for
    if $|Candidates| = 0$ then
      print ''Invalid or unrecognizable term''
      return false;
    else if $|Candidates| \geq 1$ then
      $sort\_increasing\ (Candidates, d)$
      $< c, d > \leftarrow first\_element\ (Candidates)$
      $create\_new\_rule\ (w, c, d)$
      return $< c, d >$
    end if
  else
    $< c, d > \leftarrow parse\_lemmatization\_rules\ (w)$
    return $< c, d >$
  end if
```

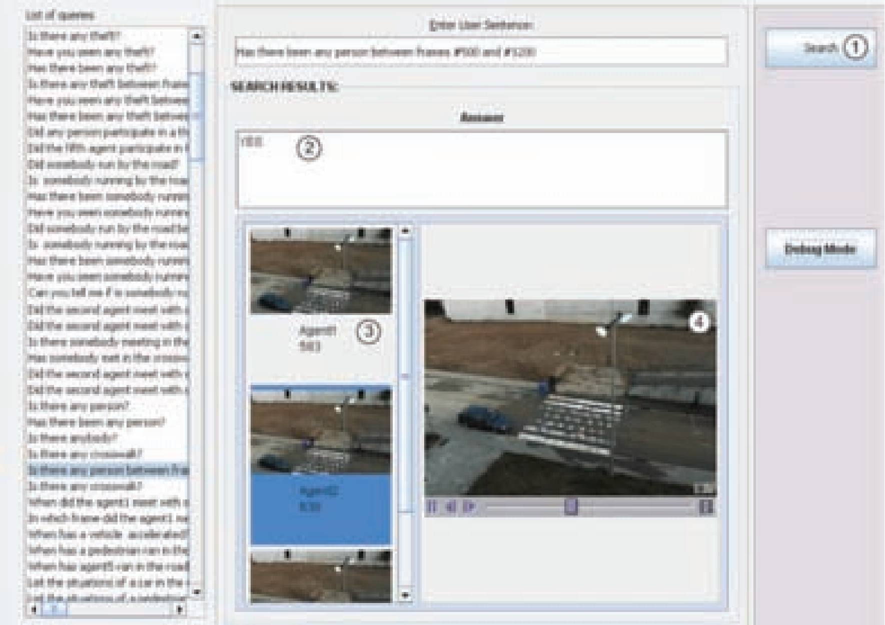
Table 5.12

Retrieve closest concept $c \subset \mathcal{T}$ to a possibly unknown word w

User query	Unknown word	Semantic distance
Have you seen any risk of runover?	*risk*	`Danger=0.0`
Has there been any crime in the scene?	*crime*	`Theft=0.18`
Show me pedestrians meeting in the pavement	*pavement*	`Sidewalk=0.0, Way=0.17, Crosswalk=0.22, Face=0.29, Road=0.29`
How many jeeps are there in the scene?	*jeeps*	`Ambulance=0.08, Car=0.08, Bus=0.08, Truck=0.17, Motorbike=0.17, Van=0.17`
How many people have picked up a backpack?	*backpack*	`Bag=0.11, Can=0.22, Bicycle=0.22, Car=0.22`
Has any pursuit happened after a theft?	*pursuit*	`Chase=0.0, Escape=0.13, Walk=0.2, Run=0.2, Zoom=0.2, Kick=0.22, Turn= 0.22, Squat=0.22, Action=0.25, Situation=0.25, Behavior=0.25, Activity=0.25`

Table 5.13

Queries including unknown words, and proposed concepts sorted by relevance.


List of queries

Is there any theft?
Have you seen any theft?
Has there been any theft?
Is there any theft between frame
Have you seen any theft betwe
Has there been any theft betwe
Did any person participate in a th
Did the fifth agent participate in t
Did somebody run by the road?
Is somebody running by the roa
Has there been somebody runnin
Have you seen somebody runnin
Did somebody run by the road be
Is somebody running by the roa
Has there been somebody runnin
Have you seen somebody runnin
Can you tell me if is somebody ru
Did the second agent meet with a
Did the second agent meet with a
Is there somebody meeting in the
Has somebody met in the crossw
Did the second agent meet with a
Did the second agent meet with a
Is there any person?
Has there been any person?
Is there anybody?
Is there any crosswalk?
Is there any person between fra
Is there any crosswalk?
When did the agent1 meet with a
In which frame did the agent1 me
When has a vehicle accelerated?
When has a pedestrian ran in the
When has agent5 ran in the road
List the situations of a car in the
List the situations of a pedestrian

Enter User Sentence:

Has there been any person between frames #500 and #1200

SEARCH RESULTS:

Answer

YES

Figure 5.17: User interface for query retrieval. When performing a query *(1)*, the system responds with a schematic textual answer *(2)*, but also with a visual list of key-frames, one for each result *(3)*. By clicking one of them, the user reproduces the video interval showing the solicited content *(4)*.

threshold distance to consider a concept has been fixed to $th = 0.20$. The average time required to solve a query has been of 1884 ± 795 ms.

Once the sentence is linked to concepts, the semantic distance of a sentence to predefined goal predicates is measured using the described Tree Edit Distance algorithm. Table 5.14 presents a list of representative queries and the system responses. These sentences have been extracted from a total amount of 110 NL queries provided by English speakers.

Finally, Fig. 5.17 shows the user interface created to facilitate query retrieval. This front-end allows users to retrieve schematic textual answers, but also to browse video responses showing the intervals where the queried contents have been observed. In addition, Fig. 5.18 depicts the rule-creation process when a concept is linked to an unknown word. The user inspects the proposed addition and can adjust the linguistic properties of the new word.

Virtual Storytelling

The objective of a Virtual Storytelling application is to automatically generate synthetic image sequences that visually explain the contents of a linguistic plot. It intends to bring high-level modeling closer to end-users, by means of a flexible solution that helps them to produce complex sequences automatically. This facilitates tasks of

User query / Goal predicate / System response
Has there been any danger of runover?
Assert{S=DangerOfRunOver}
Yes
Can you tell me whether anybody has been running by the road between frames 500 and 1500?
Assert{A=Pedestrian, S=Run, L=Road, T=(500,1500)}
Yes
When has a vehicle accelerated?
Query{T=?, A=Vehicle, S=VAccelerate}
Agent3 [852], Agent4 [939]
When has the fifth agent run by the road?
Query{T=?, A=Agent5, S=Run, L=Road}
Agent5 [1211]
How many pedestrians have entered the crosswalk?
Count{A=?, A=Pedestrian, S=Enter, L=Crosswalk}
3
What has happened in the scene after the theft?
Query{S=?, T=After(Theft)}
Chase [1241], Exit [1276]
What has happened in the scene after the theft?
Query{S=?, T=After(Theft)}
Chase [1241], Exit [1276]
What has agent2 done between frames 400 and 1100?
Query{S=?, A=agent2, T=(400,1100)}
Appear [630], Walk [642], Meet [692], Enter [799], DangerOfRunover [828], PBackUp [838], PStop [842], GiveWay [891], Enter [896], PStop [1000], LeaveObject [1054]
Who has left any object to the ground?
Query{A=?, S=LeaveObject}
Agent1 [615], Object1 [615], Agent2 [1054], Object2 [1054]
Where has agent5 gone?
Query{L=?, A=agent5}
road [1211], upper_sidewalk [1049], crosswalk [1093], upper_crosswalk [1078]
List vehicles in the scene between frames 300 and 1300
List{A=?, A=Vehicle, t=(300,1300)}
Agent3 [806], Agent4 [862]

Table 5.14

Samples of user queries. Each query instantiates a goal predicate, which in turn is transformed into a SQL query to retrieve a schematic result.

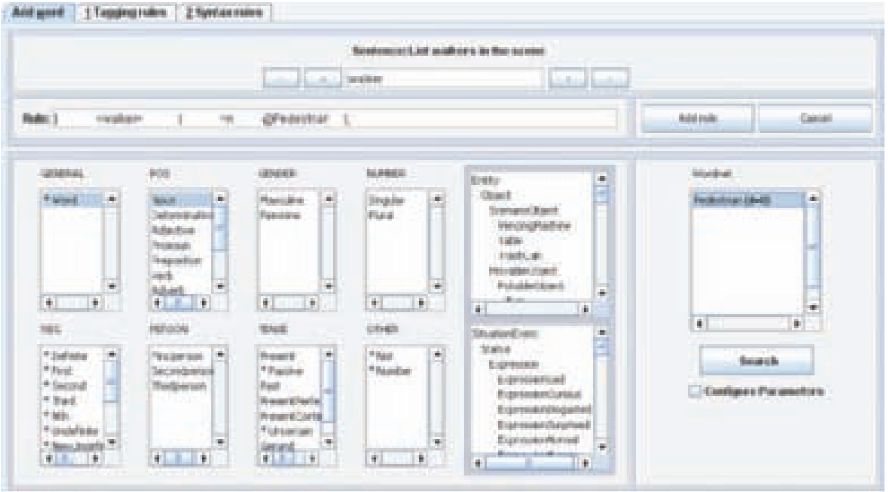


Figure 5.18: When an unknown word is lexically disambiguated using WordNet, the NLU module proposes the addition of a tagging rule to the user. The selectable linguistic features vary automatically for each language.

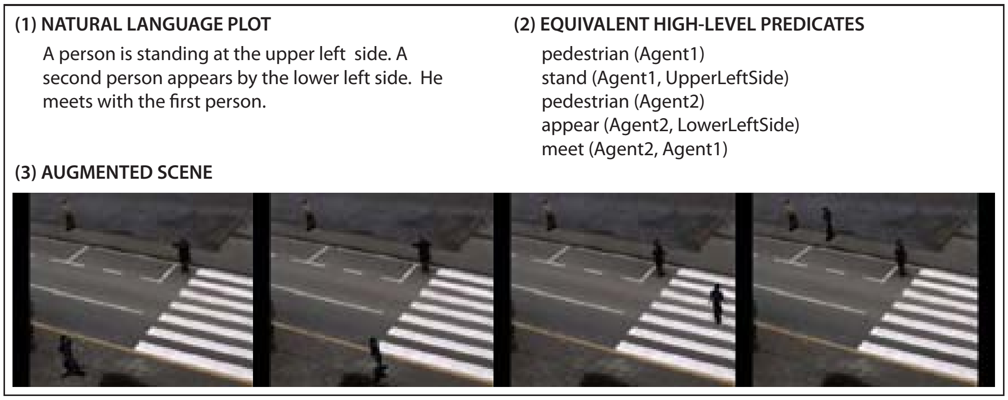
scene augmentation and simulation of agent behaviors to users of the system, and these tasks in turn enable further applications like the comparison or evaluation of tracking systems, as discussed in the next section.

Virtual storytelling requires both from the ONT–NLU and the SA modules, in order to first understand linguistic content provided by the user, and then convert this content into a visual representation of developments. The linguistic understanding of plot lines is accomplished exactly in the same way explained for query retrieval, although the goal predicates in this case are the same ones used for video reporting. Examples of this conversion are shown next.

Natural language plot	Obtained predicates
A pedestrian comes by the upper left side.	`appear(Pedestrian1,UpperLeftSide)`
Another pedestrian appears at the lower right side.	`appear(Pedestrian2,LowerRightSide)`
The first pedestrian tries to leave by the lower left side.	`leave(Pedestrian1,LowerLeftSide)`
A vehicle goes slowly by the right.	`drive(Vehicle1,RightSide,Slow)`
The second pedestrian rushes towards pedestrian 1.	`walk(Pedestrian2,Pedestrian1,Fast)`
Pedestrian 1 stops in the middle of the lower sidewalk.	`stop(Pedestrian1,LowerSidewalk)`
A new car enters by the left part.	`appear(Vehicle2,LeftSide)`
Pedestrian #2 leaves by the upper right side.	`leave(Pedestrian2,UpperRightSide)`

Each produced predicate instantiates a high-level event, which must be converted into a list of explicit spatiotemporal actions accomplished by the virtual agents. This is done by decomposing a high-level event into a temporal sequence of lower-level objectives. For instance, we may want to define a pedestrian situation *"P1 meets P2"* as the sequence (i) *"P1 reaches P2"*, and (ii) *"P1 and P2 face each other"*, or translated into FMTL predicates:

$$meet(P1, P2) \vdash go(P1, P2) \rightarrow faceTowards(P1, P2) \vee faceTowards(P2, P1) \quad (5.2)$$



(1) NATURAL LANGUAGE PLOT

A person is standing at the upper left side. A second person appears by the lower left side. He meets with the first person.

(2) EQUIVALENT HIGH-LEVEL PREDICATES

pedestrian (Agent1)
stand (Agent1, UpperLeftSide)
pedestrian (Agent2)
appear (Agent2, LowerLeftSide)
meet (Agent2, Agent1)

(3) AUGMENTED SCENE

Figure 5.19: Example of augmented scene generated from a NL textual plot. Details regarding scene augmentation are explained in following sections.

Such decompositions are modeled using SGTs, in which reaction predicates now adjust dynamically the behavior of virtual agents, instead of being `note` predicates. The generated scenes can either be completely virtual or actual augmentations of already recorded sequences. In the latter case, virtual agents can react to real occurrences, as shown in Fig. 5.19. In this example, the behavioral models encoded in SGTs establish that if the path of a pedestrian ends at the other side of the road, it must be recomputed to go through the crosswalk, and only if the crossing is granted. The concrete implementation of these tasks corresponds to the SA module, and is explained in the following sections.

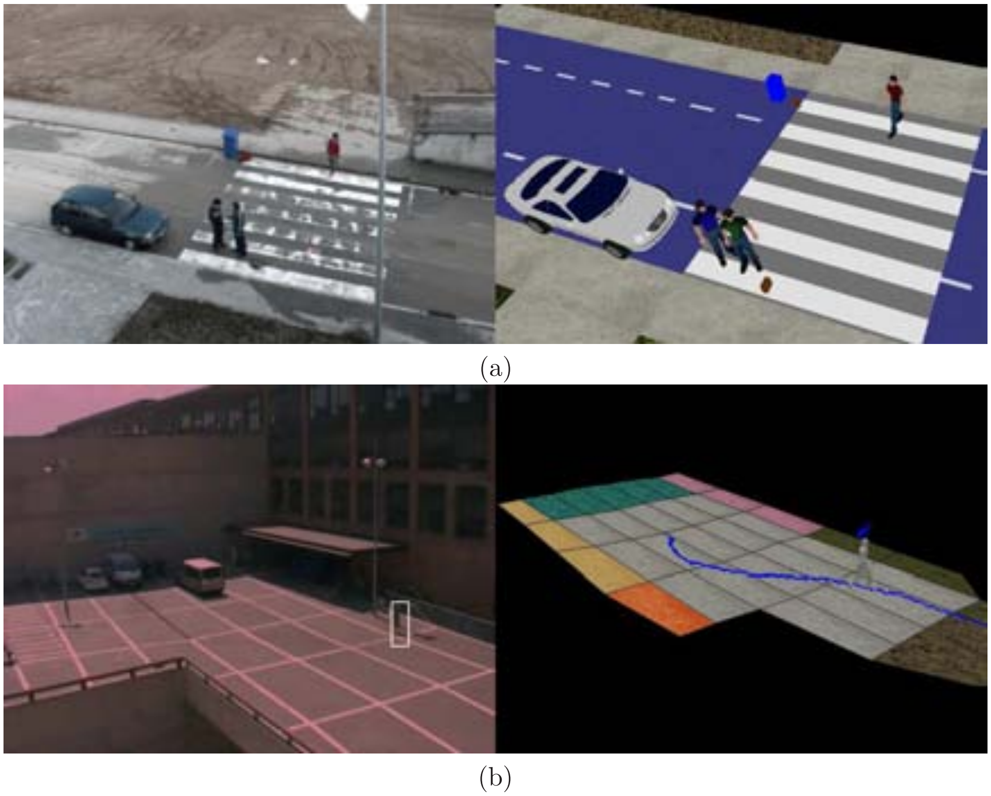
5.8 Applications of SA

Ontologies are not determinant in the case of the SA module. Nevertheless, applications like the described virtual storytelling appear from its collaborative association with the previous modules. Moreover, the addition of this module to the system enhances a multimodal interaction with end-users, by also incorporating visual languages to the communication. Three applications are considered in this section: reporting video occurrences by reconstructing them in virtual scenes; augmenting original image sequences for simulation or to test behavioral models; and application of these tasks to the evaluation of tracking systems.

Visual reporting with synthetic scenes

A completely virtual scene can be recreated from real developments observed by the system. There are several ways to achieve this, depending on the practical purpose we have, and each method entails different benefits. The implementations described here focus two main applications: (i) *visual reporting/compression/summarization* and (ii) *virtual real-time monitoring*.

In the first place, for applications of visual reporting, we base on the semantic



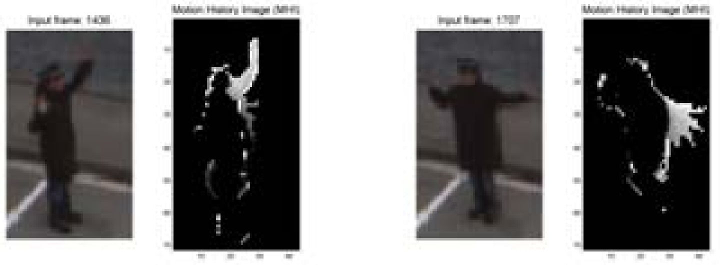
(a)

(b)

Figure 5.20: Virtual generations of the HERMES-Outdoor and ETSE-Outdoor scenes. The scene is reconstructed (a) using the list of automatically generated semantic annotations, and (b) in real-time, using instantaneous information from the trackers.

annotations obtained from the behavioral analysis detailed in Chapter 4. Only those occurrences that are relevant to the domain are considered, and the rest of visual contents are avoided. Hence, the original video sequence –8.0 Mb for 846 frames of 640×480 with high MPEG-4 compression– is converted into a list of semantic predicates –2.2 kB in plain text– that can recreate the same scene virtually, with the support of few conceptual and visual models, see Fig. 5.20(a). The main drawback in this case is the imprecision of some recreated developments, given that high-level occurrences and behaviors (such as *theft*, *chase*) are generated using predefined spatiotemporal models of action development.

On the other hand, applications of real-time monitoring and reporting do not require predefined action models, but only a rough conceptualization of the scenario. The scenario must be rich enough to let end-users understand the developments in the scene, but still limited, to avoid unnecessary delays in the processing. Fig. 5.20(b) shows an example of real-time reporting, in which the trajectories detected by the trackers are stored and used by a virtual character to recreate the scene. Additionally,



(a) (b)

Figure 5.21: The actions performed by the policeman instantiate two possible predicates: (a) `police_orders_stop(Police)` or (b) `police_orders_pass(Police)`, giving right-of-way to pedestrians and vehicles, respectively.

the numerical positions over time can be as well mapped into the corresponding semantic zones before representing the data. This would allow a major compression in expenses of a lower fidelity of the recreation.

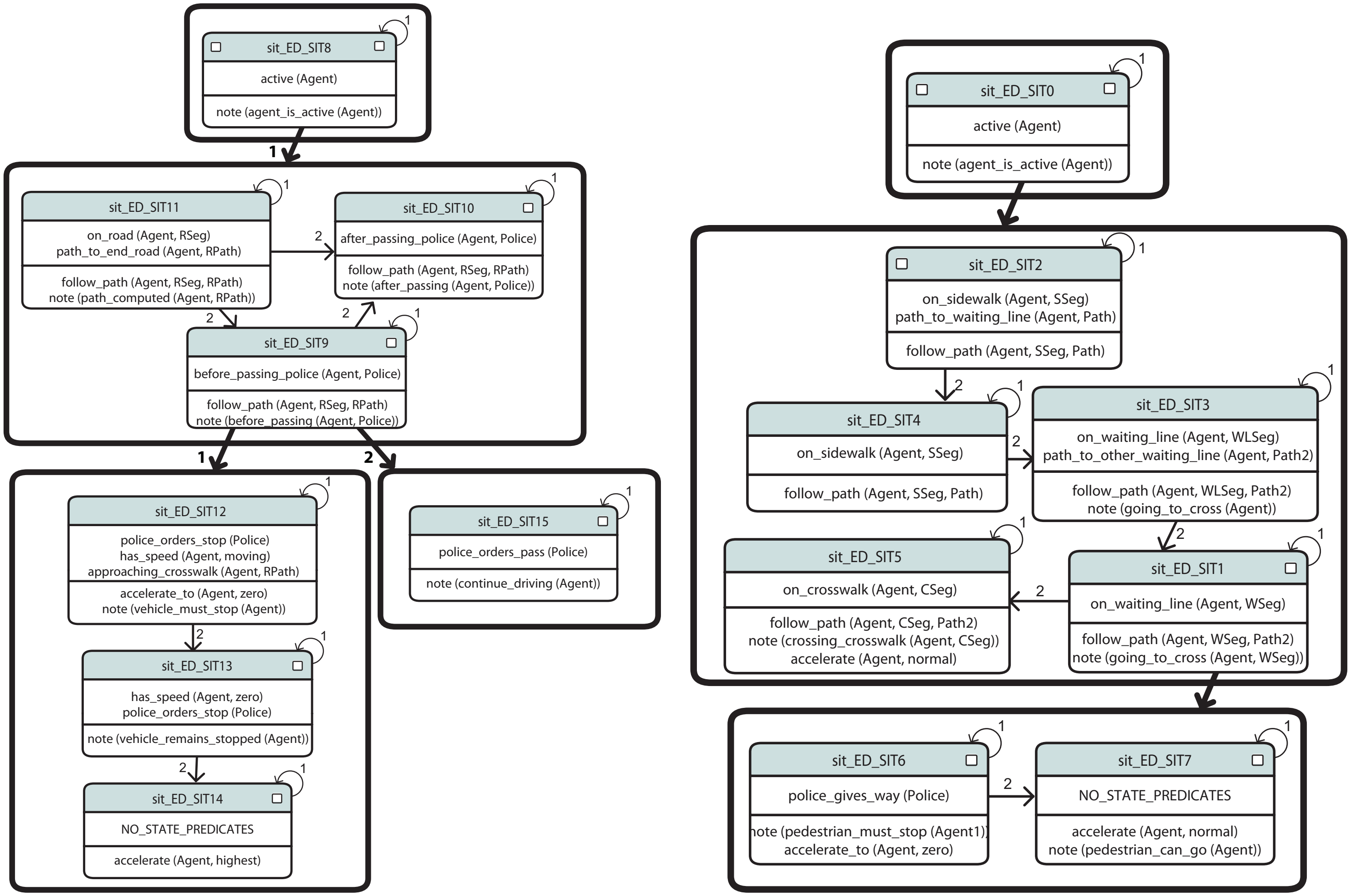
In both cases, the end-user has control over the final visualization, in terms of camera view and graphical models. Camera view control is especially beneficial for multi-camera tracking frameworks, since a proper integration of views permits end-users to overcome occlusions and have perspectives suitable to each situation.

Simulation of behaviors for autonomous agents

In this section we test the feasibility of SGTs to model synthetic behavior for virtual autonomous agents, i.e., making them reactive to (or affected by) real developments in the video sequence. In addition, this experiment considers not only tracking information at agent level, but also action recognition at body level. Hence, it also shows the flexibility to include other sources of knowledge into our behavioral framework.

In the *POLICE* sequence, a real agent acts as a policeman, giving traffic instructions to virtual agents. The policeman is tracked over time and his gestural instructions are recognized using Motion History Images [20]. This technique relates the intensity of a pixel to the temporal history of motion at that point, turning an image sequence into a monochrome image, where pixels with recent variations become brighter. Action recognition is achieved by matching the resulting images with action templates learned for different viewpoints, and generating the predicate that corresponds to the classification, see Fig. 5.21.

The predicate `police_orders_stop (Policeman)` indicates that right-of-way is given to pedestrians, and vehicles must stop. On the other hand, `police_orders_pass (Policeman)` makes pedestrians wait. Such action states are instantaneously analyzed by SGTs for both agent types, having virtual agents react to the real policeman's action following the schemes in Fig. 5.22(a) and (b). Virtual pedestrians compute a path from their initial random position to the closest waiting line of the sidewalk, and from there through the crosswalk and out of the scene. Depending on the policeman's action, pedestrians stop or not in front of the road. Similar rules apply for a virtual


sit_ED_SIT8
active (Agent)
note (agent_is_active (Agent))
sit_ED_SIT11
on_road (Agent, RSeg)
path_to_end_road (Agent, RPath)
follow_path (Agent, RSeg, RPath)
note (path_computed (Agent, RPath))
sit_ED_SIT10
after_passing_police (Agent, Police)
follow_path (Agent, RSeg, RPath)
note (after_passing (Agent, Police))
sit_ED_SIT9
before_passing_police (Agent, Police)
follow_path (Agent, RSeg, RPath)
note (before_passing (Agent, Police))
sit_ED_SIT12
police_orders_stop (Police)
has_speed (Agent, moving)
approaching_crosswalk (Agent, RPath)
accelerate_to (Agent, zero)
note (vehicle_must_stop (Agent))
sit_ED_SIT13
has_speed (Agent, zero)
police_orders_stop (Police)
note (vehicle_remains_stopped (Agent))
sit_ED_SIT14
NO_STATE_PREDICATES
accelerate (Agent, highest)
sit_ED_SIT15
police_orders_pass (Police)
note (continue_driving (Agent))
sit_ED_SIT0
active (Agent)
note (agent_is_active (Agent))
sit_ED_SIT2
on_sidewalk (Agent, SSeg)
path_to_waiting_line (Agent, Path)
follow_path (Agent, SSeg, Path)
sit_ED_SIT4
on_sidewalk (Agent, SSeg)
follow_path (Agent, SSeg, Path)
sit_ED_SIT3
on_waiting_line (Agent, WLSeg)
path_to_other_waiting_line (Agent, Path2)
follow_path (Agent, WLSeg, Path2)
note (going_to_cross (Agent))
sit_ED_SIT5
on_crosswalk (Agent, CSeg)
follow_path (Agent, CSeg, Path2)
note (crossing_crosswalk (Agent, CSeg))
accelerate (Agent, normal)
sit_ED_SIT1
on_waiting_line (Agent, WSeg)
follow_path (Agent, WSeg, Path2)
note (going_to_cross (Agent, WSeg))
sit_ED_SIT6
police_gives_way (Police)
note (pedestrian_must_stop (Agent1))
accelerate_to (Agent, zero)
sit_ED_SIT7
NO_STATE_PREDICATES
accelerate (Agent, normal)
note (pedestrian_can_go (Agent))
1
2

(a)

(b)

Figure 5.22: SGTs to constrain the behaviors of (a) virtual vehicles and (b) virtual pedestrians in the augmented sequence.

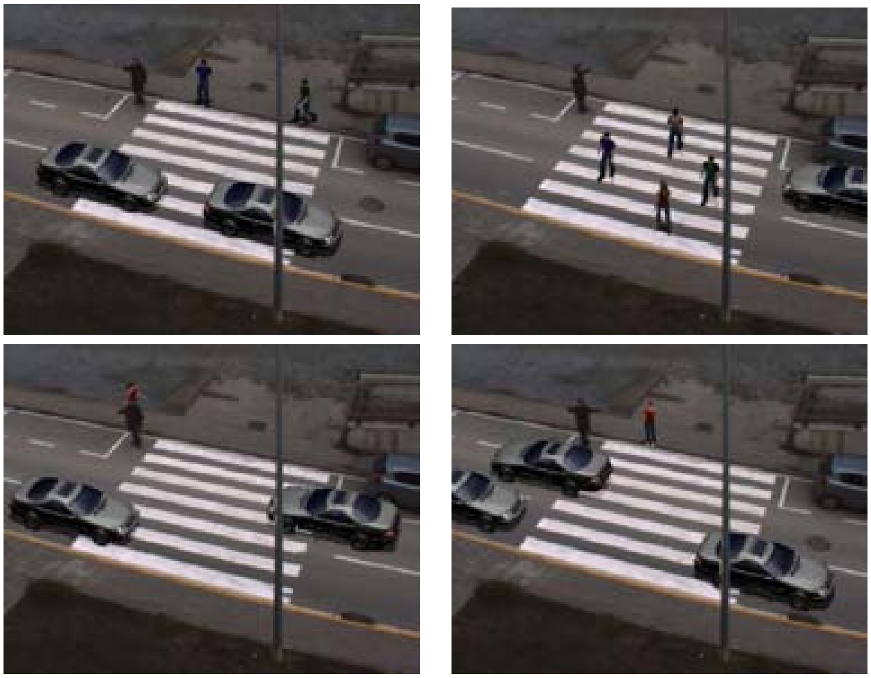


Figure 5.23: Scene augmentation of the POLICE sequence, by means of reactive virtual pedestrians and vehicles.

vehicle: if it has not passed the policeman by the road, and if the proper order is given, the vehicle stops in front of the crosswalk; otherwise it drives normally.

Fig. 5.23 shows sample frames obtained after simulating 20 virtual agents –10 vehicles and 10 humans– in the *Police* sequence. Notice that virtual agents move according to the gestures of the real policeman, and that all silhouettes are consistently maintained in the augmented sequence.

The number of virtual agents incorporated into augmented sequences affects the frame-rate of the rendering process, given the addition effort to recompute paths during the SGT traversal. We have tested the scalability of the generation of virtual agents and its consequences for a real-time performance. The experiment tests how fast the agents are generated, depending on the number of instances and the quality of the rendering, see Fig. 5.24. The code has been developed under C++ using the OpenGL library, and runs on a Pentium D 3.21 GHz, 2GB RAM. The sequences have been augmented from mid-resolution image sequences of 696 × 520 pixels. The maximum frame rate –25 fps– is achieved in most cases, and decreases as the number of agents increase.

Evaluation of trackers

This application focus on the evaluation of tracking systems specialized in open-world image sequences. State-of-the-art multi-object tracking still deals with challenges such

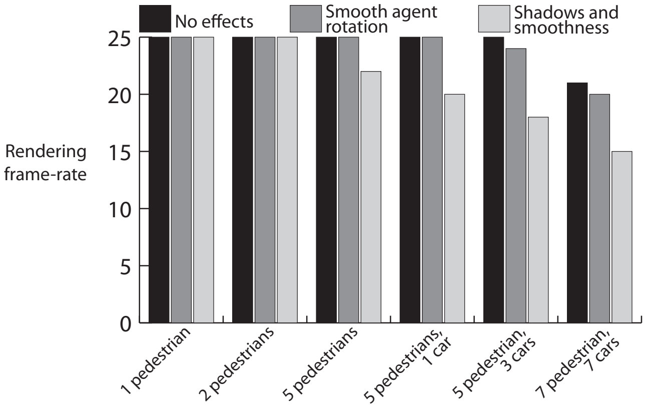


Figure 5.24: Evaluation of the rendering frame-rate by increasing the number of simultaneous agents.

as long occlusions, grouping disambiguation, or camouflage, which drive the attention of the researchers towards tools for performance evaluation and comparison. Although a high number of criteria are available to this end, a consistent evaluation of the trackers always involves to test the algorithms thoroughly over a sufficient number of sequences showing different conditions. Instead of tackling the effort-consuming task of recording new, slightly modified sequences, sometimes involving crowds of actors, it would be useful to have methods to gradually increase the difficulty of a given video sequence.

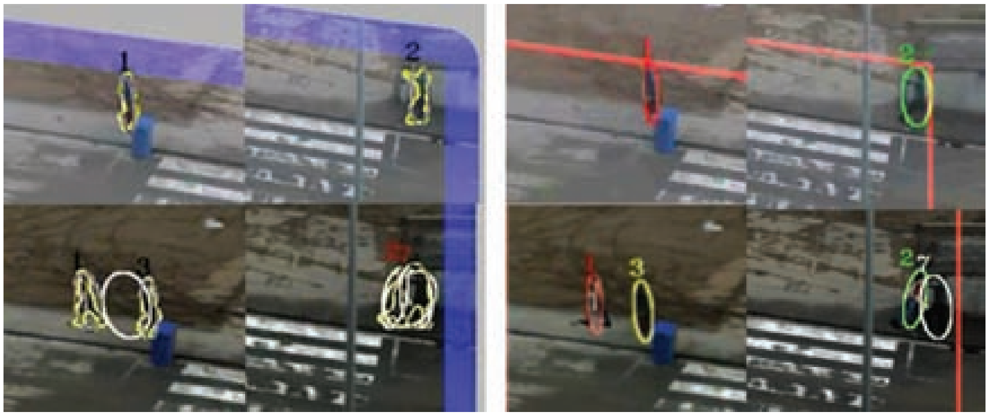
A common strategy to evaluate tracking performance is to compare the tracking results with their corresponding GT labeling. In our case, the evaluation is based on the account of basic events detectable by tracking, e.g., appearing, leaving, entering predefined semantic zones, or being occluded. The GT labeling is accomplished by manual annotation of these events, and is considered the ideal output of the trackers.

The original HERMES-Outdoor sequence has been augmented by simulating 30 new virtual agents. 15 pedestrians cross the road by the crosswalk, 10 more walk by the sidewalk, and 5 cars drive by the road in both senses. The resulting sequence has been analyzed by two trackers, a modular and hierarchically-organized tracker that switches between appearance-based and motion-based modes [108] and a real-time tracker based on segmentation by exploiting a static background [107]. A GT labeling has also been obtained manually.

The results of the evaluation are shown in Table 5.15. Due to camouflage, the number of occlusions vary substantially, although the zone-events are correctly recognized in general, with exception of few false positives. Fig 5.25 compares 2 frames showing the results of the two trackers, for the original sequence and for an augmented one. Augmentation allows us to increase the complexity of a scene in terms of involved agents. The performance of the trackers regarding the recognition of basic events can be accomplished by comparing them to a GT labeling of events.

Events	GT labeling	Hierarchical tracker	Real-time tracker
Enter scene	36	36	40
Exit scene	35	35	38
Start occlusion	17	21	17
End occlusion	17	21	15
Enter crosswalk	20	19	20
Exit crosswalk	20	19	20

Table 5.15
Evaluation of event recognition for both trackers on the augmented sequence.



(a) (b)

Figure 5.25: Scene augmentation can increase the complexity of a scene gradually, by successively adding virtual agents. Here, a hierarchical tracker (a) and a real-time tracker (b) are tested on an original sequence (top row) and its augmented equivalent (bottom row).

5.9 Discussion

This chapter has explored a series of modules enabling the communication of contents between system and end-users. Such interaction is accomplished by means of linguistic and visual interaction, and ontologically enhanced with the three modules described: ONT-NLG, ONT-NLU, and SA.

Regarding the ONT-NLG module, the ontology facilitates the structured incorporation of non-trivial knowledge to the system, such as multilingual resources for an algorithmic reporting of video contents, while allowing common processes to remain unchanged. Language extensions are easily implemented. Moreover, an ontology derives content-based summarization capabilities naturally. Further work should enhance the naturality of the produced texts, by incorporating tasks for sentence aggregation and introducing complementary words and expressions to increase expressivity.

The ONT-NLU module has proven to achieve effectively an inverse task, the algorithmic schematization of NL texts into typified predicates. Advanced interfaces for video search and browsing can be easily designed once the goal predicates are made available. Nevertheless, input queries are potentially infinite, suggesting a stronger need of recognition for the structure of the sentences. Adapting the current procedure to statistical mechanisms would be useful to add robustness to this process, something that was not necessary for generation.

Finally, scene augmentation has been demonstrated to derive applications that enhance user interaction substantially. Visual languages complement the natural ones, by offering synthetic reconstructions of observed events, or augmenting original sequences with static or dynamic elements that are controlled by the end-users. These mechanisms can be used for a variety of applications, namely simulation, evaluation of behavioral models, or comparison of tracking systems, among others.

The main limitations of this framework come from the restrictive domain of work. The linguistic models need to be extended as new situations are being detected, since the content to be communicated is provided entirely by the ontology. The chosen deterministic approach limits the variety of sentences being produced and understood, but ensures that the results will be linguistically correct, since they obey constructions proposed by native speakers. The conceptual terms on the domain can be increased or restructured by simply modifying the ontology.

Resum

Aquest capítol ha detallat tota una sèrie de mòduls pensats per a proveir comunicació entre el sistema i l'usuari final. La interacció s'acompleix mitjançant recursos de tipus visual i lingüístic, i es veu millorada ontològicament pels tres mòduls descrits: ONT-NLG, ONT-NLU i SA.

Quant al mòdul ONT-NLG, l'ontologia permet la incorporació estructurada de coneixement no trivial al sistema, com ara recursos en múltiples llengües per a la transcripció textual algorítmica de continguts de vídeo, tot i assegurant-se que els processos comuns no es canvïin. Les extensions lingüístiques s'han pogut implementar fàcilment. A més, l'ontologia incorpora naturalment la capacitat de proveir l'usuari amb resums automàtics. Futures millores en aquest mòdul s'haurien de dirigir a millorar la naturalitat dels textes generats, incorporant processos d'agregació de frases simples en compostes i en paràgrafs, i introduint paraules i expressions complementàries que enriqueixin l'expressivitat de les descripcions.

El mòdul ONT-NLU ha demostrat ser eficient en el desenvolupament de la tasca oposada, l'esquematització conceptual d'entrades de text natural fins a convertir-lo en predicats tipus. Aquest mòdul permet la creació de potents motors de cerca en bases de dades de vídeo, i l'exploració per continguts d'aquestes. No obstant, l'univers de consultes potencials és infinit, cosa que ens suggereix una necessitat més gran de fer el procés robust, quelcom que no era necessari per acomplir les funcions de generació.

Finalment, la generació d'escenes sintètiques és útil per a oferir un ample ventall d'aplicacions que milloren la capacitat d'interacció amb l'usuari de forma substancial.

Els llenguatges visuals complementen la comunicació purament lingüística amb elements estàtics i dinàmics que són fàcilment controlables pels usuaris finals. Aquests mecanismes es poden fer servir per a una varietat molt diversa d'aplicacions, des de simulació fins a avaluació de models comportamentals, passant per l'anàlisi comparatiu de sistemes de seguiment visual.

Les principals limitacions d'aquest marc de treball venen donades pel domini restringit fet servir per les aplicacions. Quan s'amplia el nombre de situacions a detectar, els models lingüístics han d'estendre's perquè el contingut a comunicar-se es basa principalment en els conceptes definits a l'ontologia. L'aproximació determinista escollida limita la varietat de frases que poden ésser generades, però assegura la correcció del text generat, donat que aquest es basa completament en construccions proposades per parlants nadius de la llengua. Els conceptes considerats al domini poden incrementar-se amb una simple modificació de l'ontologia.

Chapter 6

Concluding remarks

"One is what one is, partly at least."

Molloy (1951), by Samuel Beckett

As a conclusion to this thesis, this section revisits the main modules and contributions presented. We analyze how our ontological framework has allowed us to redirect the resources of the system to narrow gaps in different areas. The main opportunities and weaknesses of the proposed framework will be discussed, and improvements will be suggested to the problems detected, for each of the divisions of the ontological cognitive vision system.

We now revisit the main contributions for each of the main tackled fields: automatic learning, reasoning and interpretion of events and behaviors, and modules for advanced interaction. According to the distribution of gaps presented in the introduction, next table schematizes the use of a specific knowledge to solve the problems described in each chapter.

Identifier	*Specific source of knowledge*	Used in
❶	Visual representation	Chapters 3, 4
❷	Semantic / linguistic representation	Chapters 4, 5
❸	Theoretical models	Chapters 3, 4, 5
❹	User query understanding	Chapters 4, 5
❺	Communication with end-user	Chapter 5

Automatic learning

Two main tasks have been proposed for the semantic learning, namely (i) automatic labeling of semantic scenario regions, and (ii) semi-supervised incorporation of linguistic rules for NLU. The first task permits us to locate and categorize specifically a series of meaningful regions in outdoor traffic scenarios, with independence of the particular scene, and uniquely based on trajectory data and a minimal amount of ontological knowledge. The results are directly applicable to model-based reasoning tools like SGT, Petri Nets or symbolic networks, enabling them to produce richer interpretations about occurrences in a location. The second task helps end-users to progressively enrich the linguistic resources of the system, thus improving communication.

Contribution	*Knowledge implied*	*Gap to bridge*
Automatic labeling of semantic scenario regions	❶❸	Semantic
Supervised incorporation of lexical concepts by exploiting generic knowledge bases (WordNet)	❹❸❺	Model

The results of scene categorization are promising and interesting, but still not robust enough to be directly utilized for scene interpretation in every case. Right now, the technique is used in a specific domain and only exploiting motion data. By also incorporating context identification and object recognition techniques based on appearance, our method could be enhanced and extended to more complex domains –e.g., indoor scenarios, sports, social media–. Regarding NLU, new statistical techniques –e.g., those based on information theory or probabilistic parsing– could greatly improve the current performance and flexibility of the algorithms. Another promising alternative to extend the results in this field is the use of structural SVMs, which translate the accuracy of binary classifiers to environments with taxonomical organization of classes.

Reasoning and behavioral modules

One of the most important contributions of this thesis is the detailed proposal of a consistent ontological framework for cognitive surveillance. In this framework, a series of ontological resources articulate and enhance the multiple semantic processes taking place at many stages of the system. An ontology assumes the knowledge contained in the different models of the expert system –conceptual, behavioral, linguistic–, integrates them into an abstract semantic layer, and offers improved capabilities regarding their usability, interrelation, maintenance, and scalability.

We have also proposed a methodology for concept selection and top-down building and structuring of semantic models. New steps on this direction require us to investigate whether this process can be automatized with affordable risk, in which case we could reuse, merge, and grow semantic models from different domains. This task has been found to be very complex, although recent work has accomplished advances in the matter.

Contribution	*Knowledge implied*	*Gap to bridge*
Ontological framework to guide the organization and centralization of knowledge, and facilitate the maintenance and extensibility of the implied models	❸❺	Model
Detection and interpretation of semantically meaningful events from image sequences	❶❸	Semantic
Automatic indexing / annotation of video events. Content-based episodical segmentation	❶❸	Semantic
Use of high-level inferences to correct missing or corrupted sensory data	❷❸	Visual

The reasoning and interpretation modules discussed have been proven efficient to handle tasks of semantic annotation, video indexing, and content-based episodical segmentation. Nevertheless, if we take a deeper look at the current state of the reasoning system, we notice that the weight of the decisions on complex event recognition relies fundamentally on the semantic models –FMTL motion rules, SGTs–. Hence, the performance of the system is tied to the correctness of expert modeling. Although this fact provides the traditional benefits of top-down modeling paradigm that we already demonstrated, it still suggests to develop further methods that could better –i.e., more flexibly– exploit the probabilistic data retrieved from visual detectors and trackers.

One of the solutions at this regard consists of enhancing the SGT framework to take better advantage of the reasoning engine, e.g., by incorporating features such as *degrees of validity* or *multi-hypothesis inference*, which are currently not used in the situational analysis. Another possible alternative has already been suggested in Section 4.8 by means of Fuzzy Constraint Satisfaction techniques, which allow us to combine the robustness of expert systems with the flexibility and potential of current probabilistic visual detectors. This last alternative offers as an additional benefit a direct coupling with ontological resources. In any case, fuzzy techniques seem to be able to join the potentials of visual analysis and rule-based reasoning into suitable integrated solutions.

Advanced user interaction

Three different modules have been entirely designed from scratch, enabling natural and flexible interfaces to let end-users interact with the system. Natural language has been employed by two linguistic modules as a powerful tool that facilitates applications of multilingual/personalized reporting, summarization, content-based query retrieval, or storytelling for simulation. Regarding these modules for linguistic support, a natural evolution would consist of moving the communication channel from written texts to spoken dialogs, by means of Speech Recognition (SR) and Text-To-Speech (TTS) tasks. These would be attached to the current ONT-NLU and ONT-NLG modules, respectively. Both additions can build upon the already available linguistic models. An SR process, in addition, can use the restrictedness of the domain of concepts and their semantic interrelation in order to improve recognition.

Contribution	*Knowledge implied*	*Gap to bridge*
Automatic generation of reports (NL texts, synthetic animation) of relevant events in video sequences	❷❸	Interface
Summarization or compression of video information	❷❸	Interface
Multilingualism and personalization	❸	Interface
Content-based NL query retrieval	❸	Query
Camera control and update of database knowledge	❹❺	Sensory, Semantic
Simulation of behaviors via visual storytelling	❸❹❺	Semantic
Tracking performance evaluation	❸❹	Sensory

The generation and augmentation of virtual scenes has also contributed to enhance the interaction between end-users and the system. Concretely, we have proposed applications of visual reporting, simulation, compression, and performance evaluation, which complement linguistic interaction. Other effective types of user interfacing have received strong attention from the research community during the last years, such as virtual reality, haptic technologies –also applicable to virtual reality through techniques like acoustic radiation–, eye-tracking monitoring, mobile or portable devices, or multimodal interfaces combining diverse channels of interaction. Depending on the usage given to the system, the investigation of some of the techniques can suggest new trends in the accomplishment of effective and natural user interaction.

Final remarks

In this thesis we have provided a detailed framework of collaborative modules for advanced video surveillance and video understanding, based on the paradigm of HSE. A series of ontological resources have allowed us to interrelate and centralize the different types of semantic knowledge involved in the processes of generation and analysis. Furthermore, the use of ontologies enable the system to learn and organize video contents, and share them with end-users by means of advanced interfaces of communication. As a result, the ontological resources have become fundamental to narrow distinct gaps –*sensory*, *semantic*, *model*, *query*, *interface*– present in many of the tasks demanded to a cognitive vision system.

Resum

Aquesta tesi ha descrit en detall els mòduls d'alt nivell d'un sistema cognitiu artificial de visió, destinat a tasques de comprensió semàntica de seqüències de vídeo, i basat en el paradigma HSE. Una sèrie de recursos ontològics i mòduls col·laboratius ens han permès interrelacionar i centralitzar els diferents tipus d'informació semàntica involucrats en els processos de generació i anàlisi descrits. A més a més, l'ús d'ontologies ha possibilitat que el sistema extregui i organitzi de forma eficient els continguts semàntics d'un vídeo, i els comparteixi amb els usuaris finals per mitjà d'interfícies de comunicació avançadades. Com a resultat, els recursos ontològics acompleixen un paper fonamental a l'hora de superar les diferents bretxes que separen el sistema del món real i de l'usuari, els anomenats *gaps*: sensorial, semàntic, de modelat, de consulta i d'interfície. Aquests *gaps* són presents a moltes de les tasques requerides a un sistema cognitiu artificial.

Durant els diferents capítols s'han detallat tot un seguit d'aplicacions que permeten l'acompliment de d'aquestes tasques. Al capítol 3 s'ha descrit un mètode per a classificar automàticament les diferent regions semàntiques que conformen l'escenari que s'està gravant, tal com vorera, carretera, pas de vianants, zones d'espera, etcètera. Aquest mètode permet obtenir un model conceptual de l'escenari sense haver-lo de definir a mà, cosa que beneficia el posterior mòdul de reconeixement de comportament i la seva generalització a qualsevol escenari del domini.

El capítol 4 ha descrit els diferents mòduls implicats en la tasca d'interpretació de comportaments observats en seqüències de vídeo, a partir de la informació quantitativa extreta per aplicacions de seguiment visual. S'ha fet servir lògica difusa i arbres de grafs de situació per a conceptualitzar les dades, inferir nova informació i interpretar-la d'acord als models d'un domini. L'ús d'ontologies permet organitzar el coneixement d'acord amb la seva naturalesa semàntica, i fer possible futures aplicacions de recuperació de vídeos en base al seu contingut.

Finalment, el capítol 5 descriu tres mòduls que possibiliten interfícies avançades de comunicació amb l'usuari, per mitjà de la generació i comprensió de frases simples en llenguatge natural i de la generació i augmentació d'entorns virtuals. La centralització de coneixement per mitjà de l'ontologia permet reaprofitar alguns dels recursos obtinguts pel sistema (informació visual, models semàntics, consultes o respostes d'usuari) per solucionar problemes d'altres àrees, aconseguint aplicacions interessants de descripció de vídeos en múltiples llengües, cercadors i navegadors basats en llenguatge natural, resum automàtic de vídeo, o simulació i avaluació de models i tasques a partir de realitat augmentada.

Appendix A

Most frequently described events

Next we present the most frequent events detected and described by the users, see Tables A.1 and A.2. The events are sorted according to the agreement of the users to described, i.e., from most agreed –1.00 agreement means that everybody used it– to least agreed –0.00 would mean that nobody used that event–. The entity instances appearing in each fact are described in a schematic way: Ped=Pedestrian, Veh=Vehicle, Obj=Object. Shadowed facts are currently being used for automatic generation. We have used line separators to separate those events used above average, and below 10% (0.10 agreement).

Agreement	Fact
1.00	Ped1 leaves object1
1.00	Peds1,2 cross / try to cross / walk to other side / want to cross
0.90	Ped1 walks
0.86	Ped2 leaves Obj2
0.83	Ped3 runs / runs off / runs away
0.83	Peds1,2 enter crosswalk / cross / go across / go on crossing
0.83	Veh2 gives way / stops / wait for them to cross
0.80	Ped2 chases / chases after / runs after Ped3
0.70	Ped3 picks up / grabs / snatches Obj2
0.63	Peds1,2 meet / stand close
0.60	Ped3 appears / enters
0.50	Ped3 crosses
0.50	Ped3 steals / thief
0.50	Peds2 walks / comes
0.46	Ped3 walk / approaches /comes
0.46	Veh1 passes without stopping / not allowing them to cross
0.46	Veh2 appears / comes
0.43	Peds1,2 back up and stop / pull back
0.43	Peds1,2 talk / chat / have a conversation (upper crosswalk)
0.40	Ped1 stops / reaches crosswalk (ped1)
0.40	Ped2 appears
0.40	Peds1,2 stop / stand (lower crosswalk)
0.40	Veh1 appears / comes
0.36	Peds1,2 notice/realize/see Ped3
0.36	Veh1 almost hit / knock down / run over Peds1,2
0.33	Ped2,3 run
0.33	Peds1,2 shake hands (upper)
0.26	Ped1 holds briefcase / ...with a bag
0.26	Peds1,2 greet each other
0.26	Peds1,2 talk/converse/chat (lower crosswalk)
0.23	Ped1 appears
0.20	Ped1,2 keep on talking / while they talk (while crossing)
0.20	Peds1,2 stop at Veh1
0.20	Veh2 arrives / approaches at the crossing pass
0.16	object1 abandoned / forgotten
0.13	Ped2 waves / attracts attention of Ped1
0.13	Peds1,2 shake hands (lower crosswalk)
0.13	Peds1,2 still talking / keep on chatting (lower crosswalk)
0.13	Peds2,3 leave
0.13	Veh1 accelerates / goes on

Table A.1

List of events most frequently described by users (1/2). Shadowed ones are currently implemented.

Agreement	Fact
0.13	Veh1 reaches / runs towards / approaches
0.13	Veh2 exits / passes by
0.10	danger of run over / about to run over
0.10	Ped1 eventually follows the chase
0.10	Ped1 stays watching
0.10	Ped1,2 start talking (lower crosswalk)
0.10	Ped3 does not notice / ignores Obj1
0.10	Ped3 walks away from them
0.10	shout at the driver
0.10	Veh2 accelerate /drives on
0.07	Ped1 says hello to Ped2
0.07	Ped1 spins around confused / looks on bewildered / seems hesitant
0.07	Ped1 walks away
0.07	Ped2 reaches / arrives to Ped1
0.07	Ped2 tries to recover/reclaims his bag
0.07	Peds1,2 complain against / protest to car driver / raise-wave hands
0.07	Peds1,2 do not notice Ped3
0.07	Peds1,2 do not pay attention when crossing
0.07	Peds1,2 reach to the other side
0.07	Peds1,2 say goodbye to each other
0.07	Peds1,2 wait to let Veh2 pass
0.07	Veh1 leaves
0.03	brief exchange between Peds1,2
0.03	Ped1 checks road
0.03	Ped1 motions Ped2 to cross
0.03	Ped1 motions Ped2 to cross
0.03	Ped1,2 have a brief exchange
0.03	Ped1,2 out of range of vehicles
0.03	Ped2 tells Ped1 about Ped3
0.03	Ped3 bends down
0.03	Ped3 ducks
0.03	Ped3 notices Obj2
0.03	Ped3 stops near Obj2
0.03	Peds 1,2 seem to be friends
0.03	Peds1,2 are angry at Veh1
0.03	Peds1,2 are surprised
0.03	Peds1,2 communicate
0.03	Peds1,2 let the car continue its way
0.03	Peds1,2 wait for car to pass
0.03	Veh1 brakes up

Table A.2

List of events most frequently described by users (2/2). Shadowed ones are currently implemented.

Appendix B

Technical details on NL modules

This appendix aims to shed light on more technical and concrete issues which may be helpful when working with the Natural Language (NL) text generator program, in order to incorporate a new language to the already implemented ones. This part will try both to (i) structure the steps to follow and (ii) to tackle some technical issues, especially concerning the definition of parsing rules.

The architecture of the linguistic modules (Natural Language text Generation (NLG) and Natural Language text Understanding (NLU)) comes defined by two main parts: the grammars, i.e., sets of rules for forming strings in a specific natural language, and the parsers, i.e., computer processes that analyzes these rules made of sequences of tokens –typically, words–, to determine its grammatical structure with respect to the formal grammars. Both components are described next.

B.1 Grammars and metalinguistic information

The definition of formal grammars for the different natural languages requires a prior step, which is accounting the different linguistic categories and properties of a given language, which may not be the same for another one. This metalinguistic information is declared in the so-called `categories` file.

Metalinguistic information

The `categories` file lists information which is strictly related to an individual language, and characterizes it from a general point of view, for further use by the parsers. Here we can define the specific codes which will be used for each *linguistic feature* that we want to include in the NLG. We will associate a word in a language with a set of these linguistic features, e.g. *he* → *pronoun*, *masculine*, *singular*, *third person*. The list of available features will be generally different from one language to another. Each of these features will have a *tag* assigned to them in the categories file. Most of them will be employed by the different grammars to refer to a linguistic category, property, or ad-hoc identifier from the target language; at this extent, we may say that some metalinguistic information will be expressed here.

`GENERAL.Word`	`(w)`	`N`
`POS.Preposition`	`(p)`	`X`
`POS.Verb`	`(v)`	`X`
`GENDER.Masculine`	`(M)`	`X`
`NUMBER.Singular`	`(S)`	`X`
`NUMBER.Plural`	`(P)`	`X`
`REG.Definite`	`(ò)`	`N`
`REG.First`	`(ò1)`	`N`
`OTHER.Not`	`(!)`	`N`

Table B.1
Some examples from a possible *categories* file.

The syntax of each entry in this file should be the following:

```
GROUP.Feature (tag) Exclusivity
```

We can see some examples in Table B.1. Each `GROUP` includes features which refer to the same linguistic information, such as `POS`, `GENDER`, `NUMBER`, `TENSE`, etc. Referring Expression Generation (REG) group should always be included, for the onomasticon to know which tags to use when labeling the reference of an entity, since these tags will be necessary in the morphological rules.

As introduced before, a **tag** is used as a specific code to refer to a particular linguistic feature. The syntax for a tag is always the same: A unicode character optionally followed by a string of numbers. The character is case-sensitive and the length of the numeric string is undefined. There should be no repeated tags in this file (a repetition will be automatically detected by the program and logged in the error console). There exist generic tags like `w`, to refer to a word, and `!`, to express the negation of a tag, i.e. the lack of a certain linguistic feature. These tags will be maintained for a new language.

> *e.g.* Tags `REG.Definite` and `REG.First` in the example are different and thus discriminated, since the first is represented by `ò` and the second by `ò1`. To state that a word has not the REG feature `Definite`, in the rules we will use the string of two tags `!ò`.

The **exclusivity** code informs about the possibility of finding several features from the same group in a single word. An 'X' means that the feature is exclusive, and thus only one of the exclusive features from this group should be found in a single word. An 'N' means that the feature is non-exclusive and can always be added to the set of tags for the word. This was designed for validation purposes, but has not been completely implemented yet.

> *e.g.* In the given examples we find features which are exclusive, such as `Singular`. This means that a word holding this feature cannot have another exclusive one from the same group, such as `Plural`. These specifications may vary from one language to another. The non-exclusive features may be assigned to a word independently from other already included features, like in the case of `!`.

Additionally, **special groups** can also be included in these files. These are just collections of tags or lists of characters which can be identified as clusters for certain purposes. We can define special groups as needed by using this syntax:

```
=spname setOfFeatures
```

where `spname` is the name given to the special group, `setOfFeatures` is a string formed by directly appending the different characters or tags.

> e.g. In English the article *'a'* becomes *'an'* in front of a vowel. Similar phenomena occur in Spanish and Catalan, too, but in these cases we must consider the character *'h'* in addition to the vowels. Thus, we can define these special groups for English and Spanish:
>
> ```
>     vowel      aeiou  // for English
>     vowel      aeiouh // for Spanish
> ```
>
> We may also want to know whether the following word is part of the verbal phrase or not, or whether it is a non-personal conjugation of a verb (i.e. gerund, participle, infinitive). Defining special groups may help to solve the problem.

The categories file will also be in charge of containing another kind of knowledge which is inherent to the language: this is the orthographical information. Currently, this information has not been required, since the group of languages added by now share the same kind of orthographical characteristics, such as the punctuation character for end of sentence or the form in which words are separated. Incoming languages like Arabic will demand a better definition of the orthographical aspects within this file.

Grammars

In this section we discuss the different sets of linguistic rules that are required by the parsers of the generator to perform the conversion from logical predicates to NL text. First, we introduce the formalism that needs to be used to define the sets of rules. After that, we detail the particularities of each of the three grammars, viz semantic/syntactic, lexical, and morphological.

Representation formalism in Grammars: *Word* structure

The NL interface uses particular structures to represent a set of linguistic features from words in the target language. These structures contain information about the natural basic elements of a sentence, which will be appearing through the processes as individual containers of information. They are present in the grammars used for every type of analysis described in the outline, i.e. semantic/syntactical, lexical, and morphological, until the final surface form as the output of the generation system. These so-called *Word* structures are the elemental components of rules at each of these stages.

The described generation system was thought to be partially based on the Prototype Theory from Cognitive Linguistics. This theory stands upon the idea of using graded categorizations for the characterization of individuals, which are defined as collections of existent or nonexistent features. Using this approach, we ensure that:

(i) The model will be extensible. Individuals (in our case, words) which have been already defined with a set of features can be given new properties, without modifying their behaviors. This is most convenient in our case, since we deal with multiple languages.

> ***e.g.*** *Imagine that you first declare the article 'a' in English. We might say that it is a determinant (d), undefinite (U) unlike 'the', and that it is singular (S) to distinguish it from 'some', so we can apply the following tags:* $\neg dUS$*. If we add the Spanish language to the system, we then need to distinguish this word between the two genders that it may take, Masculin (M) or Feminin (F). We then have* $\langle un \rangle \neg dUMS$ *and* $\langle una \rangle \neg dUFS$*. And if we additionally include the German language, we must include also the Neutrum gender (N), so that the tags would be* $\langle ein \rangle \neg dUMS$*,* $\langle eine \rangle \neg dUFS$*, and* $\langle ein \rangle \neg dUNS$*. Still, the first English word 'a' could be identified as* $\langle a \rangle \neg dS$*.*

(ii) Different types of analysis may overlap. Sometimes it becomes impossible to perform a complete analysis assuming complete independence among the syntactic / lexical / morphological stages. Combining information from different levels of analysis can disambiguate certain situations.

> ***e.g.*** *Imagine that we must refer in English to an already mentioned entity in form of a noun phrase (NP), which is third person (N3), masculine (M), and singular (S), and that we must use a pronoun (r) for the REG expression. How to know whether we must use 'he' or 'him'? It is not enough to have the morphological characterization, but we also require the semantic valency for the entity in the sentence (the information which can be given by a syntactic/semantic structure). It is not the same to say* $NP\{S\} : \neg rN3MS \rightarrow \langle he \rangle$ *or* $NP\{DO\} : \neg rN3MS \rightarrow \langle him \rangle$*, first case for subject (S) and second for direct object (DO).*

(iii) Finally, the taxonomical organization which is indirectly included in the Prototype Theory helps to perform tasks for instantiation of individuals from an ontological point of view. This means that we can refer to subgroups of individuals assuming a certain granularity, just by combining different kinds of features from different feature

`<_>`	any word-form
`¬w`	any word
`NP:¬r`	any pronoun (defined as `r`) acting as a noun phrase (`NP`)
`NP{DO}:¬w`	any direct object (defined as `DO`) acting as a noun phrase (`NP`)
`<pre_>`	any word starting with 'pre'
`^=groupv`	any word starting with a member from a predefined group (e.g. vowels)
`be<_>¬vN3S`	any temporal form of the root verb (`v`) 'to be' which is conjugated in 3rd person (`N3`) singular (`S`)

Table B.2
EXAMPLES OF PARSING FORMULAE TARGETING DIFFERENT GROUPS OF WORDS

groups. We can refer to a specific individual or to a group of individuals accomplishing a selectable amount of characteristics, e.g. those words being in participle form, those being nouns, or those ones which are followed by a word starting with a vowel.

We define a word by several fields, following this syntax:

Cat{SynFunc}:root⟨word-form⟩¬tags

- `Cat` contains the syntactical category of the word: noun phrase, appositive, determinant...
- `SynFunc` contains the syntactical function (a.k.a. syntactical valency), such as subject, direct object, indirect object...
- `root` contains the lemma of a word. It has been mainly used for verbs, in order to express the root form, which can be very useful when dealing with irregular forms.
- `word-form` contains the word-form[1] of a word.
- `tags` contains those tags which especially refer to morphological and REG features of the word.

Nevertheless, it is generally optional to use these fields. Only the tag '$\neg w$' will be always assumed as default for a word. The other fields can be referred when necessary, using some basic grammatical conventions which enable a certain flexibility to refer specific types of words. Some examples are shown in Table B.2.

[1] A word-form is a specific production of a lexeme that contains morphological features, such as gender, number, tense, etc.

B.2 Syntax of the parsers

Next descriptions give examples of typical processes that have been required repeatedly during the implementation of the different languages. These examples explain with detail the definition of some specific rules so that they can be correctly recognized by the parsers.

Reference to *any word*

We can refer to *"any word"* by just using the expression $\neg w$. This can be used, for instance, to know whether a word is the first or last in the sentence, or to apply a rule to any word following a given expression.

```
| ¬w | <azul>¬j |;
```

e.g. $\langle verde\rangle\neg j\ \langle\ hombre\rangle\neg n\ \langle\ de\rangle\ \rightarrow \langle azul\rangle\neg j\ \langle azul\rangle\neg j\ \langle\ azul\rangle\neg j$

Verde, hombre, de → *Azul, azul, azul* (SPA)

Retrieving the n-th element from the input part of a rule

We can use the expression `$n`, where `n` is a number, to refer to the word in the `n`-th position of the input test sequence of words from the rule.

```
| AdjP:¬w NP:¬Só | Det:<a>¬dS $0 $1 |;
```

e.g. $\langle green\rangle\neg j\langle\ car\rangle\neg nN3S\acute{o} \rightarrow \langle a\rangle\neg dS\ \langle green\rangle\neg j\langle\ car\rangle\neg nN3So$

They saw green car → *They saw a green car* (ENG)

We must note that (i) the tag w is always assumed for a word, and (ii) when a REG expression is evaluated, the tag $\neg o$ (`AlreadyReferred`) replaces all REG tags.

Defining contexts

When we want to be aware of the surrounding context of a word or sequence of words, but we do not want to manipulate such a context, it is useful to define which are the edges of our 'operable region'. This can be defined by using `'['` and `']'`. The words outside these brackets will never be modified by a rule with a defined context.

```
| [ PP:<de>¬p ] <una>¬d | PP:<d'>¬p |;
```

e.g. $[\langle de\rangle\neg p]\ \langle\ una\rangle\neg dN3FS \rightarrow \langle d'\rangle\neg p$

Fent-ho tot de una → *Fent-ho tot d'una* (CAT)

Detecting beginning or ending of words

The symbol ^ is used to detect beginnings or endings. This method is useful for accomplishing contractions of words, e.g. when using apostrophes in some languages.

```
| [ PP:<de>¬p ] ^=vowel | PP:<d'>¬p  |;
```

Another possibility is to use the symbol '_' inside the word-form field. This symbol represents any string of characters. Then, $\langle\_a\rangle$ represents a word-form ending with 'a', $\langle a\_\rangle$ stands for a word-form starting with 'a', and $\langle\_\rangle$ encloses any word-form.

e.g. $[\langle de\rangle\neg p\ \langle la\rangle\neg p]\ \langle\ al\_\rangle\neg dN3FS \rightarrow \langle dell'\rangle\neg p$

De la altra strada → *Dell'altra strada* (ITA)

Note: Words united by apostrophes are currently considered separate words and contracted after morphological processing, i.e. during the implementation of orthography and surface form generation. There, $d'\ \ una$ is converted into $d'una$.

Periphrasis

When an individual grammatical concept needs to be expressed in more than one word, we can delay the expansion of the unique concept until the morphological processing. The periphrasis can be incorporated to the syntactic/semantic structure as one word using '~' as separator, and then including a morphological expansion rule like this:

```
{| PP:<pac~a>¬p | PP:<pac>¬p PP:<a>¬p |; (CAT)}
```

It is not recommended to do this process if more natural linguistics solutions are possible. This method is applicable especially for prepositional periphrasis, which contain no further linguistic interpretation individually. It should not be applied to verbal periphrasis, for instance, since this last kind often incorporates well-defined semantic separability.

B.3 Steps to implement a new language

These guidelines intend to assist the implementation of a new language for generation, by indicating which steps to follow in order to include a new language into the NL text generator at its current version. To facilitate the task, two examples for each necessary file to be created have been attached in the appendix.[2]

First of all, some general tips will be useful throughout the whole process:

[2] Note: in the files shown at the appendix, the Tag symbol for the word structure appears as ~ instead of ¬.

- Choose the most similar language implemented (if any), in order to define the new linguistic categories and grammars from existing ones. Much time can be saved by reusing information from a close language already defined.

- Be careful with the files containing rules to be parsed, especially regarding the apparition of blank lines. Only the last line from these files is blank, and there *must* be a final blank line.

Next, the main stages to cover for the implementation of the new language are enumerated:

1. Choose a 3-character identifier for the new language to implement. Add this identifier and the name of the language in the file describing the current languages in the system: `/NLInterface/common/languages.txt`, following the pattern of this file.

2. Obtain the linguistic corpus which is necessary for describing a defined set of situations in the domain of interest. Only native users can provide this information. A short, simple and natural sentence has to be expressed for each logical predicate that the system can generate.

3. Define the set of linguistic characteristics regarding metalinguistic information at the categories file, placed at `/NLInterface/common/categories_XXX.txt`, where `XXX` is the 3-character identifier for the language to implement. This includes the following steps:

   (a) Note the different linguistic features in the corpus which somehow discriminate words one from the other, and also note the roots (lemmata) of all instances found in the text.

   (b) Define the different categorical groups in the categories file, especially the PoS and REG groups. For these, one has to be aware of the different morphological categories and mechanisms of referring expression and anaphora appearing in the corpus, at least from a pragmatic point of view.

   (c) Maybe some special groups will have to be included here. Usually these are incorporated when adding morphological rules. In general, the categories file will have to be modified as needed as the implementation goes on at other stages.

4. Relate the entities and events defined in the correspondent T-Boxes (prior ontological knowledge) with the proper lemma or lemmata in the provided corpus. This means to link the different agents, locations, objects, etc. taken into account to the words or expressions found in the text. Once found, implement them in the file containing rules for lexicalization:
   `/NLInterface/rules/lexicon_XXX.txt`.

5. Divide the provided corpus into single sentences (multiple sentence aggregation is not available yet). A generic syntactic/semantic structure has to be deduced from the natural way in which each situation or idea has been expressed; each of these structures has to be linked with a particular predicate in the `/NLInterface/rules/semantics_XXX.txt` file.

6. Now a first test of the program can be run. To do so, we create a blank file which will contain the morphological files (`/NLInterface/rules/ morphology_XXX.txt`), although there is no need to fill it at the moment. A first glance at the results, with no morphological assumptions, will help to be aware of the morphological phenomena required, and will assist the design process for these rules.

7. Implement the capability of referring to previous expressions by means of anaphoras, by using the proper tags for the available REG mechanisms defined in Section 5. If the desired mechanism for REG is not present at the current system, it should be implemented in the onomasticon class, found at `/NLInterface/common/general /Onomasticon.java`. The REG tags are found mainly in the morphological rules and lexical rules files, although they can also appear at the semantic/syntactic rules file.

B.4 Referring Expression Generation

In practice, we use the codified REG tags to tell the NL text generator which REG situation to choose. These tags are mainly used to build rules for lexicalization and morphology. The generator receives a word with one or more REG tags, applies the REG operations, and replaces all REG tags by an `AlreadyReferred` one (tag ¬o in the examples), to state that the expression has been referred already. This is a control feature only, this is why it has no test functions assigned.

The REG tags are linked with REG situations in the categories file. Names of the tags can be changed freely, but not the descriptions of the situations, since they are given by the onomasticon. The only problem appears when needing a REG situation which has not been defined in the onomasticon: they will need to be coded in the proper file, as will be explained in Section B.3. On the other hand, if all REG cases for a new language have been identified previously for other languages, then no additional code needs to be added, but only the suitable descriptors for that REG case have to be used.

In the lexicalization process, we consider REG transformations within the direct lexicalization rules. When describing lexicalization, it has been said that direct mappings can only be used over agent/object entities or instances from the other classes, such as locations or directions. This is because REG expressions should apply fully over agent and object entities, whereas they usually do not apply in the same manner regarding predefined locations or directions in the scenario, i.e. circumstantial aspects. Generally there exist many ways of refer to the first-type classes, e.g. *a pedestrian*, *a new pedestrian*, *the last pedestrian*, *the first one*, *another pedestrian*, *s/he*... but references to location, time, or direction are not that rich. For instance, it is weird to use *"When I arrived home I headed to a kitchen"* if the context of the discourse suggests that there is only one instance of the location *kitchen*.

In order to refer to circumstantial aspects in this implementation, we have the possibility of either using the full specific name with a regular rule, or choosing a referring expression to avoid repetition of terms. In this last case, we should add rules replacing an instance by its class, e.g. $\langle Location \rangle$ or $\langle Direction \rangle$ instead of $\langle crosswalk \rangle$ or $\langle left \rangle$ respectively. Notice that in this case, we are incorporating the

```
\\-----------------------|--------------------------------------------------|
| (NP:<Direction>)       | (AdjP:<same>)(NP:<direction>¬ò)                 |;(ENG)
| (PP{CCPoS}:<Location>) | (PP:<in>¬p(Det:<questa>¬dFS)(NP:<posizione>¬jFS))|;(ITA)
| (PP{CCZON}:<Location>) | (PP:<en>¬p(Det:<aquesta>¬dFS)(NP:<zona>¬jFS))    |;(CAT)
\\-----------------------|--------------------------------------------------|
```

Table B.3
Lexicalization rules involving REG for repeated entities which are neither agents nor objects.

```
---------------------|----------------------------------|
| NP:¬ò              | Det:<the>¬d $0                    |;
| NP:¬nSô2            | Det:<this>¬dMS AdjP:<last>¬jMS $0 |;
| NP:¬nSõ             | Det:<a>¬d AdjP:<new>¬j $0         |;
| NPS:¬nSô            | NP:<s/he>¬rN3So                  |;
| NPDO:¬nSô           | NP:<him/her>¬rN3So               |;
| NP:¬nSô             | Det:<this>¬dMS $0                |;
| NP:¬ò1              | Det:<the>¬d AdjP:<first>¬dS $0    |;
| NP:¬ò2              | Det:<the>¬d AdjP:<second>¬dS $0   |;
---------------------|----------------------------------|
```

Table B.4
Examples showing morphological rules that involve REG tasks.

REG directly by means of these lexicalization rules.

In the morphological rules file, some REG rules have to be added to convert each instance containing REG tags into its expanded referring expression. Some examples extracted from the English morphological rules file are shown in Table B.4. The correspondence between REG tags and features is the one that has been shown on Table 5.6.

B.5 Morphological parsing

The surface realization process involves mapping the specification of a text and its constituents into a surface text form, i.e. a sequence of words, punctuation symbols, and mark-up annotations to be presented to the final user [106]. The NL text generator carries out most of this process by means of morphological parsing.

This stage probably involves the most complex grammar of the three described, but only in terms of available operations and amount of rules. The only requisites to create morphological rules are to have in mind the morphological phenomena to consider and to know the possible syntax combinations accepted by the parser, most of which have been already used in the previous sections. The morphological parser, however, allows for much flexible input and output rule syntax.

Some examples of morphological phenomena that we may need to encode are, for instance, the formation of a verb participle from its root, e.g. adding *–ed* or *–d* for regular English verbs and directly encoding the irregular ones; taking into account those words which can be affected by their neighbors, e.g. English article *a* converts into *an* in front of words starting with a vowel; and other phenomena involving contraction, change of gender or number, etc.
From the different examples described, we consider two sequential types of morphological phenomena, see Table B.5:

- First, the ones affecting single words, especially towards the generation of word-forms from the lemmata included at the lexicalization stage.

- Secondly, the ones affecting interaction of words, such as contractions or modification of word ordering. In this second type of rule, single word-forms are converted into a sequence of prosodic words[3].

The rules have to be built in a hierarchical way, so the first applicable rule is directly applied. The morphological parser has the particularity of being applied in a reiterated fashion: once a rule has been applied, the parser will search again for applicable rules from the first position (this does not apply to the previous parsers). The reason for this is that exists the possibility for a word or sequence of words to change some of its properties after the application of a rule, and hence it is possible for previously non-applicable rules to become suitable for the new morphological form.

For a more complete and extensive reference about the implemented parsers, see [36]. This document was created to propose some improvements upon the morphological stage of the Angus2 NL generator using parsing techniques, and has been the base of the currently used parsers.

At the end of the morphological process, a rich semantic/syntactic structure with referred expressions and morphological forms will be available. These structures are the ones plotted in the NL interface in form of syntactical trees. Once this structure is available, it is only necessary to perform a linearization process and to include orthographical and formatting information in order to provide the final surface form to the user.

[3] A prosodic word or phonological word is the product of the interaction between words, usually from different parts of the speech, that combine together forming a sole unit, which is not a morphological “compound word” in the generally used sense of that term[89]. For example, in the Latin sentence *“Senātus Populusque Rōmānus”* (*The Senate and the Roman People*), the word *Populusque* is formed by the noun-phrase *Populus* and the conjunctive suffix *–que*. The resulting phonological word does not coincide with a single morphological word-form.

```
  \\----------------------------|-----------------|
  | VP:<go>¬vL                  | VP:<gone>¬v     |;(ENG)
  | VP:<meet>¬vL                | VP:<met>¬v      |;(ENG)
  | VP:<_>¬vL                   | VP:<_ed>¬v      |;(ENG)
  \\----------------------------|-----------------|
| PP:<a>¬p Det:<el>¬dMS       | PP:<al>¬pdMS    |;(CAT/SPA)
  | PP:<per>¬p Det:<el>¬dMS     | PP:<pel>¬pdMS   |;(CAT)
  | [ Det:<_>¬dS ] ^=vowel      | Det:<l’>¬d      |;(CAT)
  | [ PP:<de>¬p ] ^=vowel       | PP:<d’>¬p       |;(CAT)
  | [ Det:<quest_>¬dS ] ^=vowel | Det:<quest’>¬dS |;(ITA)
  \\----------------------------|-----------------|
```

Table B.5

Examples of some simple morphological rules in Catalan, English, and Italian. The upper ones, in English, allow to obtain the participle (tag $\neg L$) of a verb ($\neg v$). The third rule is general, the two first are examples of exceptions and shall appear first. In the second set of rules, the Catalan and Italian ones, prosodic manipulation is allowed. The two first examples of this second set enable contractions of certain prepositions and determiners; the three last examples show the situation in which certain words appearing in front of a word starting by vowel experiment apostrophication.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases*. Addison Wesley Publi. Co., London, 1995. [Page **57**]

[2] J.K. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999. [Page **23**]

[3] M. Al-Hames and G. Rigoll. A multi-modal mixed-state dynamic Bayesian network for robust meeting event recognition from disturbed data. In *IEEE International Conference on Multimedia and Expo (ICME 2005).*, pages 45–48, 2005. [Pages **30, 31** and **84**]

[4] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea. A constrained probabilistic Petri Net framework for human activity detection in video. *IEEE Transactions on Multimedia*, 10(6):982–996, October 2008. [Pages **31** and **33**]

[5] M. Arens, R. Gerber, and H.H. Nagel. Conceptual representations between video signals and natural language descriptions. *Image and Vision Computing*, 26(1):53–66, 2008. [Pages **8** and **35**]

[6] M. Arens and H.-H. Nagel. Behavioral knowledge representation for the understanding and creation of video sequences. In *Proc. of the 26th German Conference on Artificial Intelligence (KI'2003)*, pages 149–163, September 2003. [Page **63**]

[7] M. Arens, A. Ottlik, and H.-H. Nagel. Natural language texts for a cognitive vision system. In *Proc. of the 15th European Conference on Artificial Intelligence (ECAI'2002)*, pages 455–459, July 2002. [Page **8**]

[8] F. Baader, D. Calvanese, D. McGuiness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic handbook*. Cambridge University Press, Cambridge, UK, 2003. [Pages **36** and **57**]

[9] F. Baader, D. Calvanese, D.L. McGuinness, P. Patel-Schneider, and D. Nardi. *The description logic handbook: theory, implementation, and applications*. Cambridge Univ Pr, 2003. [Pages **7** and **36**]

[10] P. Baiget, C. Fernández, X. Roca, and J. Gonzàlez. Automatic learning of conceptual knowledge for the interpretation of human behavior in video sequences. In *Proc. of the 3rd IbPRIA*, volume 4477, pages 507–514, Girona, Spain, 2007. Springer LNCS. [Page **68**]

[11] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The berkeley framenet project. In *Proc. of the COLING-ACL*, Montreal, Canada, 1998. [Page **35**]

[12] S. Balcisoy, M. Kallman, R. Torre, P. Fua, and D. Thalmann. Interaction techniques with virtual humans in mixed environments. pages 205–216, 2001. [Page **36**]

[13] S. Balcisoy and D. Thalmann. Interaction between real and virtual humans in augmented reality. *Computer Animation*, pages 31–38, Jun 1997. [Pages **35** and **36**]

[14] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Effective codebooks for human action recognition. In *Proc. of ICCV Workshop on Video Oriented Event Categorization*, 2009. [Page **84**]

[15] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video event classification using string kernels. *Multimedia Tools and Applications*, 48:69–87, 2010. [Page **33**]

[16] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *CVPR*, pages 1–8, Anchorage, USA, 2008. [Pages **27** and **28**]

[17] M. Bertini, A. Del Bimbo, and G. Serra. Learning rules for semantic video event annotation. In *Proceedings of the international conference on Visual Information Systems (VISUAL)*, 2008. [Pages **37** and **62**]

[18] P. Bille. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239, 2005. [Page **110**]

[19] J. Black, D. Makris, and T. Ellis. Hierarchical database for a multi-camera surveillance system. *Pattern Analysis and Applications*, 7(4):430–446, 2004. [Page **45**]

[20] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. In *Royal Society workshop on knowledge-based vision*, volume B-352, pages 1257–1265, 1997. [Pages **24** and **131**]

[21] F. Bobillo and U. Straccia. fuzzyDL: An expressive Fuzzy Description Logic reasoner. In *Proc of the Int. Conf. on Fuzzy Systems (FUZZ-08)*, 2008. [Page **83**]

[22] K. Bontcheva. Generating tailored textual summaries from ontologies. In *Proc. of the Extended Semantic Web Conference*, 2005. [Page **36**]

[23] B. Bose and E. Grimson. Improving object classification in far-field video. In *CVPR*, volume 2, 2004. [Page **42**]

[24] F. Brémond. *Scene understanding: Perception, multi-sensor fusion, spatio-temporal reasoning and activity recognition.* PhD thesis, HDR Université de Nice-Sophia Antipolis, Nice Cedex, France, 2007. [Page **24**]

[25] G.J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *10th European Conference on Computer Vision, Part I*, page 44. Springer, 2008. [Page **29**]

[26] H. Buxton. Generative models for learning and understanding dynamic scene activity. In *ECCV workshop on generative model-based vision*, Copenhagen, Denmark, June 2002. [Page **23**]

[27] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence magazine*, 78(1-2):431–459, 1995. [Page **35**]

[28] P. Cimiano, U. Reyle, and J. Saric. Ontology driven discourse analysis for information extraction. *Data and Knowledge Engineering Journal*, 55:59–83, 2005. [Page **37**]

[29] W. Croft and DA Cruse. *Cognitive linguistics.* Cambridge Univ Press, 2004. [Page **42**]

[30] H.M. Dee, R. Fraile, D.C. Hogg, and A.G. Cohn. Modelling scenes using the activity within them. In *Proc. of the International Conference on Spatial Cognition VI: learning, reasoning, and talking about space*, page 408. Springer, 2008. [Pages **29** and **39**]

[31] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. of VSPETS*, 2005. [Page **83**]

[32] M. Douze and V. Charvillat. Real-time generation of augmented video sequences by background tracking. *Computer Animation and Virtual Worlds*, 17(5):537–550, 2006. [Page **35**]

[33] A. Ekin and A.M. Tekalp. Generic event detection in sports video using cinematic features. In *Computer Vision and Pattern Recognition Workshop, 2003*, volume 4, 2003. [Page **27**]

[34] C. Fellbaum. *WordNet: an electronic lexical database.* MIT press, Massachusetts Institute of Technology. Cambridge, Massachusetts., 1998. [Pages **35**, **107** and **110**]

[35] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006. [Pages **41**, **42** and **44**]

[36] C. Fernández. Addition of a post-processing stage in the surface realization module of a nlg. Technological report, Computer Vision Center, Bellaterra, Spain, March 2007. [Page **159**]

[37] C. Fernández, P. Baiget, F.X. Roca, and J. Gonzàlez. Interpretation of complex situations in a cognitive surveillance framework. *Signal Processing: Image Communication*, 23(7):554–569, August 2008. [Pages **102** and **110**]

[38] A. Fernandez-Caballero, F.J. Gomez, and J. Lopez-Lopez. Road-traffic monitoring by knowledge-driven static and dynamic image analysis. *Expert Systems with Applications*, 35(3):701–719, October 2008. [Page **32**]

[39] GL Foresti, L. Marcenaro, and CS Regazzoni. Automatic detection and indexing of video-event shots for surveillance applications. *IEEE Transactions on Multimedia*, 4(4):459–471, 2002. [Page **9**]

[40] F. Fusier, V. Valentin, F. Brémond, M. Thonnat, M. Borg, D. Thirde, and J. Ferryman. Video understanding for complex activity recognition. *Machine Vision and Applications*, 18(3):167–188, 2007. [Pages **27, 31** and **33**]

[41] E. Gelenbe, K. Hussain, and V. Kaptan. Simulating autonomous agents in augmented reality. *Journal of Systems and Software*, 74(3):255–268, 2005. [Page **35**]

[42] R. Gerber and H.-H. Nagel. (Mis-?)Using DRT for generation of natural language text from image sequences. In *Proc. ECCV'98*, volume 2, pages 255–270, Freiburg, Germany, 1998. LNCS 1407. [Page **93**]

[43] J. Gonzàlez, D. Rowe, J. Varona, and X. Roca. Understanding dynamic scenes based on human sequence evaluation. *Image and Vision Computing*, 27(10):1433–1444, 2009. [Pages **16, 27, 32, 33, 63, 71** and **83**]

[44] Jordi Gonzàlez. *Human sequence evaluation: the key-frame approach*. PhD thesis, Universitat Autònoma de Barcelona, 2004. [Pages **24, 115** and **117**]

[45] G. Granlund. *Cognitive vision systems. Organization of architectures for cognitive vision systems*, pages 37–55. Springer Verlag, 2006. [Page **34**]

[46] J.M. Gryn, R.P. Wildes, and J.K. Tsotsos. Detecting motion patterns via direction maps with application to surveillance. *Computer Vision and Image Understanding*, 113(2):291–307, 2009. [Page **28**]

[47] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43:625–640, November/December 1995. [Pages **8** and **58**]

[48] M. Haag, W. Theilmann, K. Schäfer, and H.-H. Nagel. Integration of image sequence evaluation and fuzzy metric temporal logic programming. In *Proc. of the 21st annual German conference on AI (KI 97)*, pages 301–312, London, UK, 1997. Springer-Verlag. [Page **68**]

[49] A. Hartholt, T. Russ, D. Traum, E. Hovy, and S. Robinson. A common ground for virtual humans: using an ontology in a natural language oriented virtual human architecture. In *Language Resources and Evaluation Conference (LREC)*, 2008. [Page **37**]

[50] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Press, 2003. [Page **41**]

[51] J. Hartz, L. Hotz, B. Neumann, and K. Terzic. Automatic incremental model learning for scene interpretation. In *Proceedings of the International Conference on Computational Intelligence (IASTED CI-2009)*, Honolulu, USA, 2009. [Page **29**]

[52] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *International Conference on Computer Vision*, pages 84–93, 2001. [Pages **24**, **30**, **31**, **32** and **33**]

[53] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer. Video content annotation using visual analysis and large semantic knowledgebase. In *Proceedings of the conference on Computer Vision and Pattern Recognition*, 2003. [Page **35**]

[54] E.H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence magazine*, 63(1-2):341–385, 1993. [Page **34**]

[55] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE transactions on systems, man, and cybernetics*, 34:334–352, 2004. [Page **23**]

[56] W. Hu, D. Xie, and T. Tan. A hierarchical self-organizing approach for learning the patterns of motion trajectories. *IEEE Transactions on Neural Networks*, 15(1):135–144, 2004. [Pages **27** and **28**]

[57] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, and Dan Xie. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, 2006. [Pages **27** and **28**]

[58] N. Ikizler and D.A. Forsyth. Searching video for complex activities with finite state models. In *CVPR*, 2007. [Pages **31** and **33**]

[59] A. Jaimes and S. Chang. A conceptual framework for indexing visual information at multiple levels. In *Proceedings of the IS&T SPIE Internet Imaging*, 2000. [Page **36**]

[60] Neil Johnson and David Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC*, pages 583–592, Surrey, UK, UK, 1995. BMVA Press. [Page **27**]

[61] H. Kamp and U. Reyle. *From discourse to logic*, volume I, II. Kluwer Academic Publishers, Dordrecht, Boston, London, 1993. [Pages **7** and **93**]

[62] H. Kamp and U. Reyle, editors. *Semantics of some temporal expressions. How we say WHEN it happens. Contributions to the theory of temporal reference in natural language*. Max Niemeyer Verlag, Tuebingen, Germany, 2001. [Page **95**]

[63] H. Kamp, J. van Genabith, and U. Reyle. *Discourse representation theory*, chapter 3, pages 21–128. Kluwer Academic Publishers, 2004. [Page **93**]

[64] K.M. Kitani, Y. Sato, and A. Sugimoto. Recovering the basic structure of human activities from noisy video-based symbol strings. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(8):1621–1646, 2008. [Pages **30** and **33**]

[65] P. Kohli, L. Ladickỳ, and P.H.S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. [Page **29**]

[66] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002. [Page **34**]

[67] I. Kompatsiaris, Y. Avrithis, P. Hobson, and M.G. Strinzis. Integrating knowledge, semantics and content for user-centred intelligent media services: the acemedia project. In *Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'04)*, pages 21–23, Lisboa, Portugal, 2004. [Page **35**]

[68] Yiannis Kompatsiaris and Paola Hobson, editors. *Semantic multimedia and ontologies: theory and applications*. Springer, 2008. [Pages **24**, **32** and **36**]

[69] M.P. Kumar, P.H.S. Torr, and A. Zisserman. Obj cut. In *CVPR*, 2005. [Pages **29** and **40**]

[70] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. *Advances in Neural Information Processing Systems*, 16:1–8, 2004. [Pages **29** and **40**]

[71] G. Lakoff. *Women, fire, and dangerous things*. University of Chicago Press, 1987. [Page **92**]

[72] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. [Page **46**]

[73] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. of CVPR*, 2008. [Page **83**]

[74] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *IEEE TSMC, Part C*, 39(5):489–504, September 2009. [Pages **7**, **8**, **23**, **32**, **33** and **54**]

[75] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8, 2007. [Pages **30** and **33**]

[76] T.L. Le, A. Boucher, M. Thonnat, and F. Brémond. A framework for surveillance video indexing and retrieval. In *International Workshop on Content-Based Multimedia Indexing (CBMI 2008)*, pages 338–345, 2008. [Page **62**]

[77] Y. Lee, S. Oh, and W. Woo. A context-based storytelling with a Responsive Multimedia System (RMS). In *Proc. of the 3rd International Conference on Virtual Storytelling: using virtual reality technologies for storytelling*. Springer, 2005. [Page **36**]

[78] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*, pages 383–395. Springer, 2008. [Page **28**]

[79] S.Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag, 2001. [Pages **30, 40** and **41**]

[80] J. Lou, Q. Liu, T. Tan, and W. Hu. Semantic interpretation of object activities in a surveillance system. In *International Conference on Pattern Recognition*, volume 16, pages 777–780, 2002. [Page **34**]

[81] M. Ma and P. Mc Kevitt. Visual semantics and ontology of eventive verbs. In *Proc. of the 1st International Joint Conference on Natural Language Processing*, pages 278–285, 2004. [Pages **35, 36** and **56**]

[82] N. Magnenat-Thalmann and D. Thalmann. Virtual humans: thirty years of research, what next? *The Visual Computer*, 21(12):997–1015, 2005. [Page **35**]

[83] D. Mahajan, N. Kwatra, S. Jain, P. Kalra, and S. Banerjee. A framework for activity recognition and detection of unusual activities. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*. Citeseer, 2004. [Pages **31** and **33**]

[84] N. Maillot, M. Thonnat, and A. Boucher. Towards ontology-based cognitive vision. *Machine Vision and Applications*, 16(1):33–40, 2004. [Pages **27** and **36**]

[85] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE TSMC, Part B*, 35(3):397–408, June 2005. [Pages **27** and **28**]

[86] H. Marburger, B. Neumann, and H.J. Novak. Natural language dialogue about moving objects in an automatically analyzed traffic scene. In *Proc. 7th IJCAI, Vancouver*, pages 49–51, 1981. [Page **34**]

[87] J. Martí, J. Freixenet, J Batlle, and A. Casals. A new approach to outdoor scene description based on learning and top-down segmentation. *Image and Vision Computing*, 19:1041–1055, 2001. [Page **54**]

[88] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, 2003. [Pages **30** and **33**]

[89] P. Matthews. *Morphology*. Cambridge University Press, New York, USA, 2nd edition edition, 1991. ISBN 0-521-41043-6 (hb). ISBN 0-521-42256-6 (pbk). [Pages **7** and **159**]

[90] S.J. McKenna and H.N. Charif. Summarising contextual activity and detecting unusual inactivity in a supportive home environment. *Pattern Analysis and Applications*, 7(4):386–401, 2004. [Pages **27** and **28**]

[91] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *Proceedings of the National Conference on Artificial Intelligence*, pages 770–776. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002. [Pages **30**, **31** and **33**]

[92] B. Morris and M. Trivedi. Learning trajectory patterns by clustering: experimental studies and comparative evaluation. In *CVPR*, 2009. [Page **27**]

[93] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6:59–74, 1988. [Page **24**]

[94] H.-H. Nagel. Steps towards a cognitive vision system. *Artificial Intelligence magazine*, 25(2):31–50, 2004. [Pages **7** and **32**]

[95] H.-H. Nagel and R. Gerber. Representation of occurrences for road vehicle traffic. *Artificial Intelligence magazine*, 172(4–5):351–391, 2008. [Pages **7**, **27**, **31**, **33**, **62**, **67** and **70**]

[96] R. Nevatia, J. Hobbs, and B. Bolles. An ontology for video event representation. In *Proceedings of the international workshop on Detection and Recognition of Events in Video*, 2004. [Pages **36** and **84**]

[97] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. [Pages **30** and **33**]

[98] F. Nilsson. *Intelligent network video: understanding modern video surveillance systems*. CRC Press, 2009. [Page **14**]

[99] S. Nirenburg and V. Raskin. *Ontological semantics*. MIT Press Boston, MA, 2004. [Page **61**]

[100] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:831, 2000. [Pages **30**, **31** and **84**]

[101] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: a survey. *Artificial Intelligence for Human Computing*, pages 47–71, 2007. [Pages **23** and **34**]

[102] G. Papagiannakis, S. Schertenleib, B. O'Kennedy, M. Arevalo-Poizat, N. Magnenat-Thalmann, A. Stoddart, D. Thalmann, S. Geneva, and S. Lausanne. Mixing virtual and real scenes in the site of ancient Pompeii. *Computer Animation and Virtual Worlds*, 16(1):11–24, 2005. [Page **36**]

[103] S. Park and J.K. Aggarwal. Event semantics in two-person interactions. In *Proc. of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 4, pages 227–230, Washington, DC, USA, 2004. IEEE Computer Society. [Pages **24**, **30**, **35** and **84**]

[104] C. Piciarelli and G. L. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15):1835–1842, 2006. [Pages **27** and **28**]

[105] F. Porikli and T. Haga. Event detection by eigenvector decomposition using object and frame features. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 114–114, 2004. [Pages **27** and **28**]

[106] E. Reiter and R. Dale. *Building natural language generation systems*. Cambridge University Press, Cambridge, UK, 2000. [Pages **90**, **92**, **97**, **104**, **105** and **158**]

[107] D. Roth, E. Koller-Meier, and L. Van Gool. Multi-object tracking evaluated on sparse events. *Multimedia Tools and Applications*, pages 1–19, September 2009 online. [Pages **30**, **42** and **134**]

[108] D. Rowe, I. Rius, J. Gonzàlez, and J.J. Villanueva. Improving tracking by handling occlusions. In *3rd ICAPR*, volume 2, pages 384–393, UK, 2005. Springer LNCS. [Pages **30** and **134**]

[109] Daniel A. Rowe. *Towards robust multiple-target tracking in unconstrained human-populated environments*. PhD thesis, Universitat Autònoma de Barcelona, 2008. [Page **71**]

[110] C. Saathoff and S. Staab. Exploiting spatial context in image region labelling using fuzzy constraint reasoning, 2008. Last access Nov. 2009. [Page **83**]

[111] A.M. Sanchez, M.A. Patricio, J. Garcia, and J.M. Molina. A context model and reasoning system to improve object tracking in complex scenarios. *Expert Systems with Applications*, 36(8):10995–11005, October 2009. [Page **32**]

[112] K. Schäfer and C. Brzoska. "F-Limette": fuzzy logic programming integrating metric temporal extensions. *Journal of Symbolic Computation*, 22:725–727, 1996. [Pages **7** and **68**]

[113] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. of ICPR*, 2004. [Page **83**]

[114] J. Shotton, M. Johnson, R. Cipolla, T.C.R.D. Center, and J. Kawasaki. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, 2008. [Pages **29** and **40**]

[115] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009. [Page **29**]

[116] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000. [Pages **16** and **53**]

[117] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006. [Page **33**]

[118] P. Smith, N. da Vitoria, and M. Shah. Temporal boost for event recognition. In *10th IEEE International Conference on Computer Vision*, October 2005. [Pages **30** and **33**]

[119] S. Staab and R. Studer. *Handbook on ontologies*. Springer, 2004. [Page **36**]

[120] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000. [Pages **27** and **28**]

[121] P. Sturgess, K. Alahari, P.H.S. Torr, and UK Oxford. Combining appearance and structure from motion features for road scene understanding. In *British Machine Vision Conference*, 2009. [Page **29**]

[122] L. Talmy. *Toward a cognitive semantics – Vol. 1: Concept structuring systems*. Bradford Book, 2000. [Page **58**]

[123] C. Town. Ontological inference for image and video analysis. *Machine Vision and Applications*, 17(2):94–115, 2006. [Page **36**]

[124] R. Troncy, O. Celma, S. Little, R. Garcıa, and C. Tsinaraki. Mpeg-7 based multimedia ontologies: Interoperability support or interoperability issue. In *1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies*, pages 2–15, 2007. [Page **36**]

[125] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: a survey. *IEEE Transactions on Circuits, Systems, and Video Technologies*, 18:1473–1488, 2008. [Page **23**]

[126] K. Vadakkeveedu, P. Xu, R. Fernandes, and R.J. Mayer. A content based video retrieval method for surveillance and forensic applications. In *Proceedings of SPIE*, volume 6560, page 656004, 2007. [Page **37**]

[127] J. Van Benthem and A.G.B. ter Meulen. *Handbook of logic and language*. Elsevier, North Holland, 1997. [Page **8**]

[128] R. Vezzani and R. Cucchiara. Visor: Video surveillance on-line repository for annotation retrieval. In *International Conference on Multimedia and Expo*, pages 1281–1284, Hannover, Germany, June 2008. IEEE Computer Society. [Page **32**]

[129] M. Vincze, W. Ponweiser, and M. Zillich. Contextual coordination in a cognitive vision system for symbolic activity interpretation. In *Proc. of the 4th IEEE International Conference on Computer Vision Systems*, volume 1, pages 12–12, Washington DC, USA, 2006. IEEE Computer Society. [Pages **33, 34** and **35**]

[130] V.T. Vu, F. Brémond, and M. Thonnat. Automatic video interpretation: A recognition algorithm for temporal scenarios based on pre-compiled scenario models. *Computer Vision Systems*, pages 523–533, 2003. [Pages **31** and **33**]

[131] X. Wang, K.T. Ma, G.W. Ng, and W.E.L. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *CVPR*, pages 1–8, 2008. [Pages **27** and **28**]

[132] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV*, pages 110–123, Graz, Austria, 2006. [Pages **27** and **28**]

[133] G. Wilcock. Talking OWLs: towards an ontology verbalizer. In *Proc. of the International Semantic Web Conference*, 2003. [Page **36**]

[134] R.A. Wilson and F.C. Keil, editors. *The MIT encyclopedia of the cognitive sciences*. Bradford Book, 2001. [Page **91**]

[135] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, pages 37–44, 2006. [Pages **29** and **40**]

[136] T. Xiang and S. Gong. Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006. [Pages **30** and **33**]

[137] Z. Xiong, X.S. Zhou, Q. Tian, Y. Rui, and H. TS. Semantic retrieval of video-review of research on video retrieval in meetings, movies and broadcast news, and sports. *IEEE Signal Processing Magazine*, 23(2):18–27, 2006. [Page **33**]

[138] B. Yao, L. Wang, and S. Zhu. Learning a scene contextual model for tracking and abnormality detection. In *CVPR workshops*, pages 1–8, 2008. [Pages **27** and **28**]

[139] G. Zhang, X. Qin, X. An, W. Chen, and H. Bao. As-consistent-as-possible compositing of virtual objects and video sequences. *Computer Animation and Virtual Worlds*, 17(3-4):305–314, 2006. [Page **35**]

[140] T. Zhang, H. Lu, and S. Li. Learning semantic scene models by object classification and trajectory clustering. In *CVPR*, 2009. [Pages **27** and **28**]

[141] Z. Zhang, K. Huang, T. Tan, and L. Wang. Trajectory series analysis based event rule induction for visual surveillance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. [Page **27**]

[142] Z. Zhang and J.A. Miller. Ontology query languages for the semantic web: A performance evaluation. *Journal of Web Semantics*, 2005. [Page **36**]

[143] H. Zheng, H Wang, and N. Black. Human activity detection in smart home environment with self-adaptive neural networks. In *Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC)*, pages 1505 –1510, april 2008. [Pages **30** and **33**]

Publications

Refereed journals

- Carles Fernández, Pau Baiget, Xavier Roca, Jordi Gonzàlez. Enhancing the Semantic Content of Already Recorded Surveillance Sequences. *Pattern Recognition Letters (Accepted with changes)*. Elsevier.

- Carles Fernández, Pau Baiget, Xavier Roca, Jordi Gonzàlez. Determining the Best Suited Semantic Events for Cognitive Surveillance. Submitted to *Expert Systems with Applications (Under review process)*. Elsevier.

- Carles Fernández, Pau Baiget, Xavier Roca, Jordi Gonzàlez. Interpretation of Complex Situations in a Cognitive Surveillance Framework. *Signal Processing: Image Communication Journal*, Special issue on 'Semantic Analysis for Interactive Multimedia Services'. Elsevier, volume 23, issue 7, pp. 554–569, August 2008.

- Carles Fernández, Xavier Roca, Jordi Gonzàlez, Providing Automatic Multilingual Text Generation to Artificial Cognitive Systems. *Vigo International Journal of Applied Linguistics*, number 5, pages 37–62, October 2008.

- Pau Baiget, Carles Fernández, Xavier Roca, Jordi Gonzàlez. Generation of augmented video sequences combining behavioral animation and multi–object tracking. *Computer Animation and Virtual Worlds.* Volume 20, Issue 4, Pages 447–489, July/August 2009.

Book chapters

- Carles Fernández, Pau Baiget, Xavi Roca, Jordi Gonzàlez. Exploiting Natural Language Generation in Scene Interpretation. Chapter 4 of *Human-Centric Interfaces for Ambient Intelligence*, pages 71–93. Elsevier Science and Technology Book Group, October 2009.

Refereed major conferences

- Nicola Bellotto, Eric Sommerlade, Ben Benfold, Charles Bibby, Ian Reid, Daniel Roth, Luc Van Gool, Carles Fernández, Jordi Gonzàlez. A Distributed Camera System for Multi-Resolution Surveillance. In *3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC2009).* Como, Italy, September 2009.

- Carles Fernández, Pau Baiget, Jordi Gonzàlez. Mixed-Initiative Authoring for Augmented Scene Modeling. In *22nd Annual Conference on Computer Animation and Social Agents (CASA 2009).* Amsterdam, The Netherlands, June 2009.

- Pau Baiget, Carles Fernández, Xavier Roca, Jordi Gonzàlez. Autonomous Virtual Agents for Performance Evaluation of Tracking Algorithms. In *5th International Conference on Articulated Motion and Deformable Objects (AMDO'2008).* **Best paper award**. Andratx, Spain, July 2008.

- Carles Fernández, Jordi Gonzàlez. Ontology for Semantic Integration in a Cognitive Surveillance System. In *2nd international conference on Semantics And digital Media Technologies (SAMT'2007)*, Genova, Italy, December 2007.

- Carles Fernández, Pau Baiget, Xavier Roca, Jordi Gonzàlez. Natural Language Descriptions of Human Behavior from Video Sequences. In *30th Annual German Conference on Artificial Intelligence (KI-2007).* Osnabrück, Germany, October 2007.

- Carles Fernández, Pau Baiget, Xavier Roca, Jordi Gonzàlez. Semantic Annotation of Complex Human Scenes for Multimedia Surveillance. In *10th International Conference on Advances in AI (AI*IA 2007).* Roma, Italy, September 2007.

- Pau Baiget, Carles Fernández, Xavier Roca, Jordi Gonzàlez. Automatic Learning of Conceptual Knowledge for the Interpretation of Human Behavior in Video Sequences. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007).* Girona, Spain, June 2007.

Other conferences and workshops

- Carles Fernández, Pau Baiget, F. Xavier Roca, Jordi Gonzàlez. Cognitive-Guided Semantic Exploitation in Video Surveillance Interfaces. In *Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS Workshop)*, in conjunction with *British Machine Vision Conference (BMVC'2008).* Leeds, UK, September 2008.

- Pau Baiget, Carles Fernández, Xavier Roca, Jordi Gonzàlez. Observing Human Behavior in Image Sequences: the Video-Hermeneutics Challenge. 3rd CVC Workshop: Progress of Research and Development (CVCRD'2008). Cerdanyola del Vallès, Barcelona, Spain, October 2008.

- Carles Fernández, Pau Baiget, F. Xavier Roca, Jordi Gonzàlez. Three Dialogue-based Challenges for Cognitive Vision Surveillance. 3rd CVC Workshop: Progress of Research and Development (CVCRD'2008). Cerdanyola del Vallès, Barcelona, Spain, October 2008.

- Pau Baiget, Carles Fernández, Aariel Amato, F. Xavier Roca, Jordi Gonzàlez. Constructing a Path Database for Scene Categorization. In *2nd CVC Workshop: Progress of Research and Development (CVCRD'2007).* Cerdanyola del Vallès, Barcelona, Spain, October 2007.

- Carles Fernández, Pau Baiget, F. Xavier Roca, Jordi Gonzàlez. High-level Integration for Cognitive Vision Surveillance. In *2nd CVC Workshop: Progress of Research and Development (CVCRD'2007).* Cerdanyola del Vallès, Barcelona, Spain, October 2007.

- Carles Fernández, Pau Baiget, Mikhail Mozerov, Jordi Gonzàlez. Spanish Text Generation for Human Evaluation using FMTHL and DRS. In *1st CVC Workshop on the Progress of Research and Development (CVCRD 2006).* Cerdanyola del Vallès, Barcelona, Spain, October 2006.

- Pau Baiget, Carles Fernández, Xavier Roca, Jordi Gonzàlez. Interpretation of Human Motion in Image Sequences Using Situation Graph Trees. In *First CVC Workshop on the Progress of Research and Development (CVCRD 2006).* Cerdanyola del Vallès, Barcelona, Spain, October 2006.

Technical Reports

- Carles Fernández, Jordi Gonzàlez, A Multilingually-Extensible Module for Natural Language Generatio. CVC Technical Report 120, UAB, January 2008.

- Carles Fernández , Natural Language for Human Behavior Evaluation in Video Sequences, CVC Technical Report 101, UAB, February 2007.