Interpretation of Complex Situations in a Semantic-Based Surveillance Framework

Carles Fernández [a,*], Pau Baiget [a], Xavier Roca [a],

Jordi Gonzàlez [b]

[a] *Computer Vision Centre, Edifici O. Campus UAB, 08193, Bellaterra, Spain*

[b] *Institut de Robòtica i Informàtica Ind. UPC–CSIC, 08028, Barcelona, Spain*

Abstract

The integration of cognitive capabilities in computer vision systems requires both to enable high semantic expressiveness and to deal with high computational costs as large amounts of data are involved in the analysis. This contribution describes a cognitive vision system conceived to automatically provide high level interpretations of complex real-time situations in outdoor and indoor scenarios, and to eventually maintain communication with casual end users in multiple languages. The main contributions are (i) the design of an integrative multilevel architecture for cognitive surveillance purposes; (ii) the proposal of a coherent taxonomy of knowledge to guide the process of interpretation, which leads to the conception of a situation-based ontology; (iii) the use of situational analysis for content detection and a progressive interpretation of semantically rich scenes, by managing incomplete or uncertain knowledge, and (iv) the use of such an ontological background to enable multilingual capabilities and advanced end-user interfaces. Experimental results are provided to show the feasibility of the proposed approach.

Key words: Cognitive vision system, situation analysis, applied ontologies

1 Introduction

A notable evolution has occurred in the fields of artificial intelligence and cognitive science during the last years [27]. Symbolic and connectionist approaches, traditionally antagonistic views of intelligence, are currently sharing common ground and seen as complementary for the development of hybrid intelligent systems. Especially in modern cognitive science, interdisciplinarity plays a fundamental role in the design of cognitive models, capable of reasoning and adapting under strong constraints of uncertainty.

Cognitive systems, unlike traditional intelligent machines, do not pursuit reasoning as an end in itself, nor try to design generalized models or absolute truth. Instead, they highlight the need to use situated frameworks to enable actions which are desirable in concrete, natural contexts and toward specific goals. These systems incorporate plausible computational mechanisms which approximate human-like cognitive operations of perception, reasoning, decision, learning, reaction, or communication, in order to enhance the human capacity to recognize and interpret meaningful content in large collections of information acquired from diverse sources.

We aim to design the integrative architecture of a *Cognitive Vision System* (CVS), which extracts descriptions of interpreted human behaviors and complex situations from recorded video sequences. We focus on controlled scenarios

* Corresponding author.
Email address: perno@cvc.uab.es (Carles Fernández).

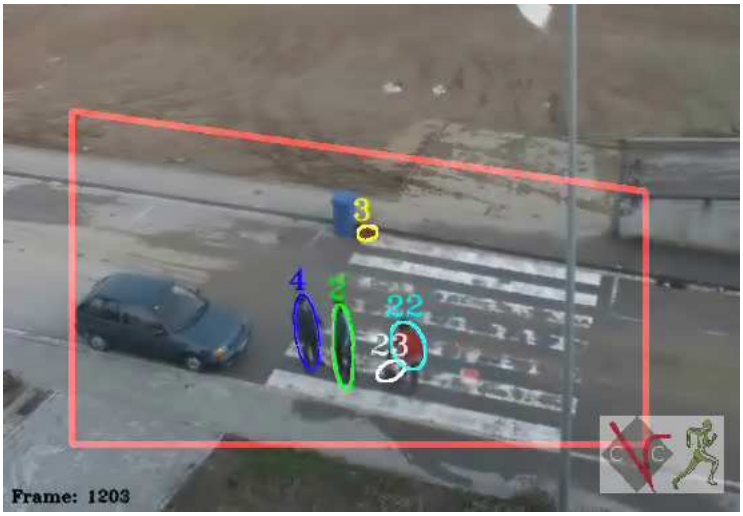


Fig. 1. Snapshot from the recordings on one scenario covered by the system. In this scene, a pedestrian (Agent 22) is stealing an object (Object 23) from another pedestrian (Agent 2).

of a restricted discourse domain, such as pedestrians interacting with vehicles and static objects in inner-city scenarios, see Fig. 1. The CVS is thought to have multilingual capabilities and an interface with several modalities of communication with external users, especially involving Natural Language (NL) interaction.

Such characteristics demand a series of considerations to fulfil regarding the integration of different types of knowledge: symbolic and non-symbolic representations; episodic and semantic content; metric-temporal and linguistic-oriented predicates. This article discusses the different modules implied in the semantic levels of the considered system, describing the alternatives which have been chosen to clarify the structure of the domain knowledge, and to devise effective cooperation between the representation formalisms appearing.

This contribution enhances current research on semantics in multimedia services by proposing a model of high-level architecture, which enables the in-

tegration of semantic information, its interpretation in situated domains, and the interaction with external end-users. This has been particularly applied to the analysis of information received from multiple cameras (active and passive) placed in the considered scenarios. Such an architecture is open to the addition of new sources, e.g. audio information.

The paper is organized as follows: next section reviews some work related to the field of this project. Section 3 structures the CVS by proposing a comprehensive multimodal architecture. Following chapters are structured according to the semantic expressivity reached by the modules described in them: First, Section 4 introduces the basic terminology of concepts selected for the system and compiled into an ontology, which is presented as a convenient tool for semantic integration. Different representational formalisms are chosen as suitable for the different types of knowledge implied in the classification. Section 5 discusses the processes implied at the conceptual level, which carry out tasks for conceptualization, reasoning, and inference over the quantitative knowledge obtained from the visual subsystem. Section 6 describes how higher interpretations are obtained, as a result of a situational analysis at the behavioral level using contextual and intentional models. After that, Section 7 gives a general description of the user interaction level, which is designed to provide a complete interface of communication with a final user of the system. Section 8 shows experimental results for the main stages discussed, and Section 9 concludes the article by highlighting some important considerations and pointing out future lines of work.

2 Related Work

The system conceived in this article especially builds upon the work done by Nagel, who has actively investigated for decades the field of CVS applied to vehicular traffic surveillance [19,2]. He tackles the high-level analysis of visual occurrences by means of fuzzy logic inference engines, and derives the results to the generation of NL textual descriptions. An enhancement of this original scheme has been developed by Gonzàlez, which enables to enlarge the domain of a CVS towards the analysis of general human behaviors in image sequences, in what has been called Human Sequence Evaluation (HSE) [10]. Our system builds upon this framework.

The evaluation of human behaviors in image sequences is a commonly required task for applications such as surveillance, content-based retrieval of documents, or advanced interfaces related to cognitive fields. Nevertheless, while low-level visual techniques have been actively investigated for decades, high-level processing has acquired significant attention in the field of vision especially in the last years, as stated in [21]. The automatic analysis and description of temporal events was tackled many years ago by Marburger et al. [17], who proposed a NL dialogue in German to retrieve information about traffic scenes. More recent methods to identify human activities in video sequences have been reported by Kojima et al. [14], which are based in concept hierarchies of actions to recognize interesting elements and developments in a scene, particularly interactions between people and objects. In [7], Buxton reviews progress in generative models for advanced CVS to explain activities in dynamic scenes, observing applications such as education, smart rooms, and also surveillance systems. The study of interactions among moving people and

objects is undertaken under statistical approaches for high-level attention and control. Most approaches do not emphasize the episodic properties of analyzed behaviors; instead, we define an independent stage to analyze the evolution of situations and their contextualization. The integration of asserted facts to interpret a certain situation is made using a fuzzy metric-temporal engine, which takes both geometric and temporal considerations into account.

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 [15]. Crowley proposes some conceptual frameworks towards the understanding of observed human activity, including interactions [8]. He suggests a model consisting of a set of roles to be accomplished by entities; specific configurations of interrelated entities playing roles conform the so-called situations. 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 [26]. 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 [16]. Several methods for categorizing eventive verbs are discussed, and the notion of visual valency is introduced as a semantic modeling tool. In [21], 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 [6], 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. However, our requirement for multilingual capabilities demands a neutral semantic stage which holds language independency; this has been solved by defining a concept hierarchy which is not based on eventive verbs, but on the generality of situations, that allows structuring the domain knowledge from a more comprehensive point of view.

Current CVSs build 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 [26,11]. 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.

3 An Architecture for CVS

The cognitive system presented in this article performs HSE and is built upon three disciplines, namely computer vision, knowledge representation, and computational linguistics. It follows the multilevel architecture for human-oriented evaluation of video sequences proposed in [10], known as *Human Sequence Evaluation* (HSE). Its modular scheme suggests a bidirectional flow of communication between consecutive layers, in order to:

(1) Evaluate the information at a certain cognitive stage, and generate new descriptions oriented to modules at higher levels (bottom-up data flow). The produced results come after the analysis of knowledge validated by lower levels, in combination with predefined goal-oriented models and the history of asserted facts up to the moment.
(2) Send high-level productions and inferences back in a reactive manner, to support low-level processing and guide mechanisms for evaluation (top-down data flow).

This composed configuration results in an active chain of cooperative interactions among the different modules for visual, conceptual, and linguistic processing, see Fig. 2. Following the HSE scheme, we define several levels of sensors and actuators: in the first place, the *Active Sensor Level* (ASL) consists of a distributed layout of static and active Pan–Tilt–Zoom cameras which constantly provide the vision system with image sequences from distinct viewpoints. At the *Image Signal Level* (ISL), a segmentation process is performed, which allows to detect significant information within the image data. Three levels of granularity are considered for the detection of agent trajectories, body

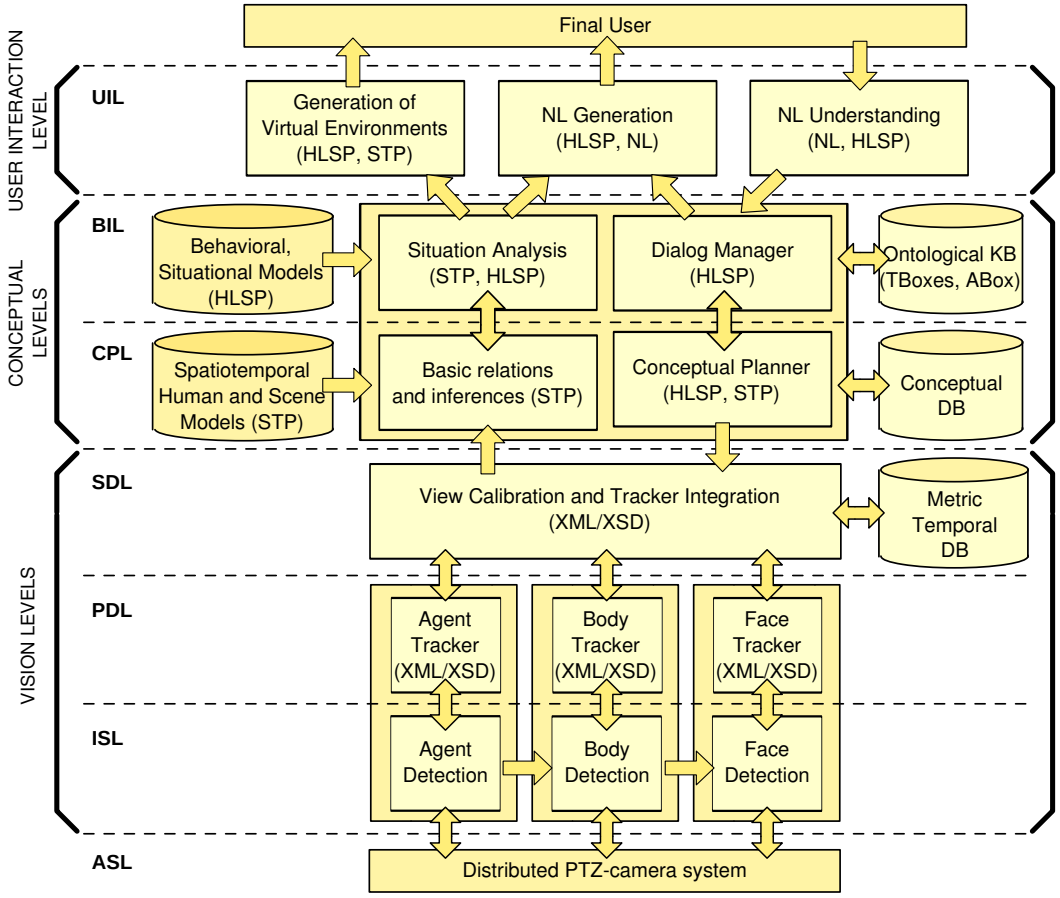


Fig. 2. General architecture for the considered system. The knowledge representation formalism used by each module is enclosed in parenthesis. A deeper analysis over these formalisms is done throughout the following sections.

postures, and facial expressions. Next, image features resulting from the segmentation process are manipulated by means of tracking procedures at the *Picture Domain Level* (PDL), which maintains the identification of targets. The information about detected features is forwarded at each time-step to the *Scene Domain Level* (SDL), in which a supervisor process takes control over the trackers and integrates data into a unique database of quantitative information. At this point, the geometric data for each detected agent over time is available within a ground-plane representation of the controlled scenario.

The syntactical interoperability among the modules is accomplished by using XML for data exchange purposes. In addition, XML schemas have been selected to provide content control over the communication, yet granting the scalability and flexibility of this process. On the other hand, the capabilities for semantic interoperability have been introduced into the system architecture by means of an ontology, which restricts the validity and relationships of the involved semantic terms to the specific working domain. Next section describes this idea in more detail.

We proceed on the basis that visual and structural information consisting of geometrical values are available over time. Details about the extraction of visual information that we use can be found in [13] and [23]. Next sections will deal in more detail with the high-level stages of the system, namely *Conceptual Predicates* (CPL), *Behavior Interpretation* (BIL), and *User Interaction* (UIL) levels.

4 Knowledge Representation

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 [16], 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.

The design of the ontology for the described CVS has been done putting especial effort on the definition of the knowledge base. Description Logic allows us to structure the domain of interest by means of *concepts*, designing sets of objects, and *roles*, denoting binary relations between concept instances [3]. 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/Situation-TBox (see Table 1), Entity-TBox, and Descriptor-TBox (see Table 2).
- 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)`

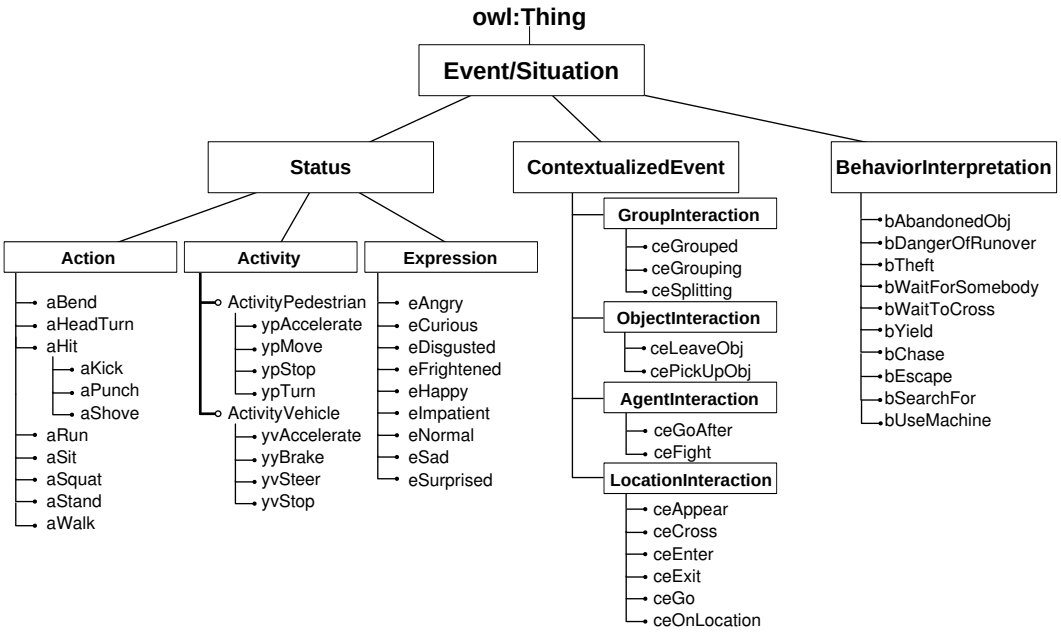


Table 1

Taxonomy containing some concepts from the Event/Situation-TBox.

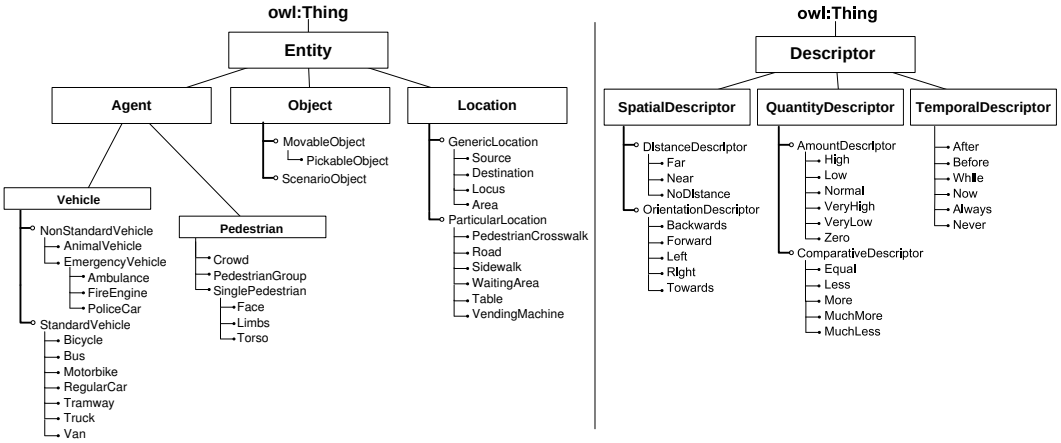


Table 2

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

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].

	Description	Examples for: Agent	Body	Face
(Pose Vectors)	Array of Static Configurations over time	trajectory of locations	sequence of postures, gestures	sequence of expressions
Status	Dynamic interpretaton of static elements	*stopped: sitting? / standing?*	*standing, moving head up and down*	*serious expr.: worry? / concentration?*
Contextualized Event	Disambiguation using information from other levels	*standing in front of a scene object*	*nod? / search for something?*	*staring at scene object*
Interpreted Behavior	Hypothesis of interpretation using complete scene knowledge	*"A person is searching for information on the timetables of the subway station"*		

Table 3

A knowledge-based classification of human motion representations. This terminology, which structures the Event/Situation-TBox, will guide the process of interpretation.

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 [25]. For him, semantic understanding involves the combination of these domains into an integrated whole. Our classification of situations (i.e. the Event/Situation-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 3:

- First, *Pose Vectors* are collections of detected static configurations for the tracked elements, such as positions or orientations at a given time-step. 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 represent 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 CVS –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.

5 Conceptual Reasoning

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 [12]:

(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.

Fuzzy Metric Temporal *Horn* Logic (FMTHL) has been conceived as a suitable mechanism to solve each of the aforementioned demands [24]. It is a rule-based inference engine in which conventional logic formalisms are extended

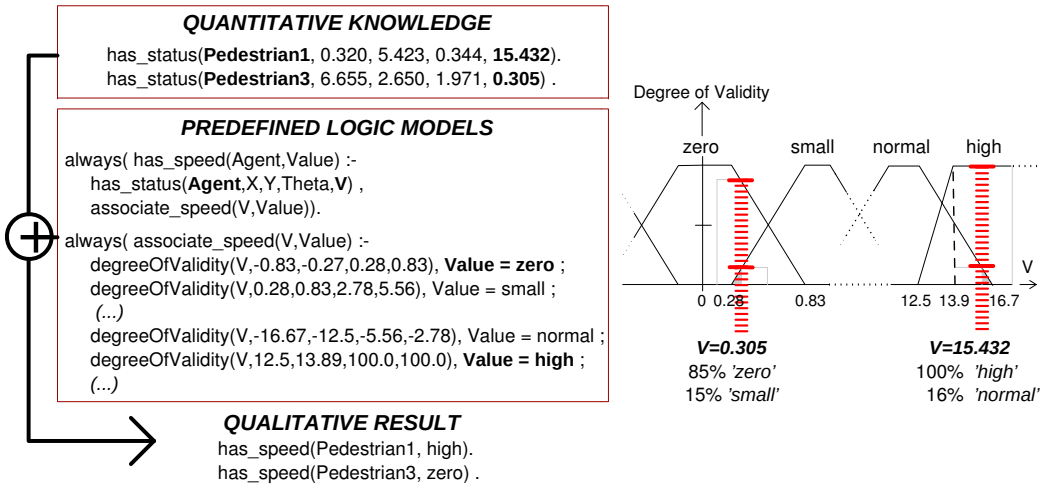
by a temporal and a fuzzy component. This last one enables to cope with uncertain or partial information, by allowing variables to have degrees of truth or falsehood. The temporal component permits to represent and reason about propositions qualified in terms of time. These propositions are represented 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 [4]:

$$t \quad ! \quad \texttt{has\_status (Agent, X, Y, Theta, V)} \tag{1}$$

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 (*Theta*) 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 Location-TBox are assigned to them:

```
    point (14, 5, p42)
   line (p42, p43, l42)
segment (l31, l42, lseg_31)
crosswalk_segment (lseg_31)
```
(2)



(a) (b)

Fig. 3. Conversion from quantitative to qualitative values. (a) The input status vectors contain information for the agents. As a result, qualitative descriptions are represented logically. (b) FMTHL includes fuzzy mechanisms accepting more than one single interpretation, since it confers *degrees of validity* to values on uncertain ranges.

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. 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 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. 3 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 [18]. 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 *Spatiotemporal Predicates* (STP). More specifically, STP facilitate a schematic representation of knowledge which is time-indexed and incorporates uncertainty. Hence, all those concepts in the Event/Situation-TBox which can be inferred only using these constraints are enclosed under this category. STP do not consist only of atomic predicates, but they can be produced after an interpretation based on temporal-geometric considerations. Next example shows FMTHL inference rules for the STP `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),
```

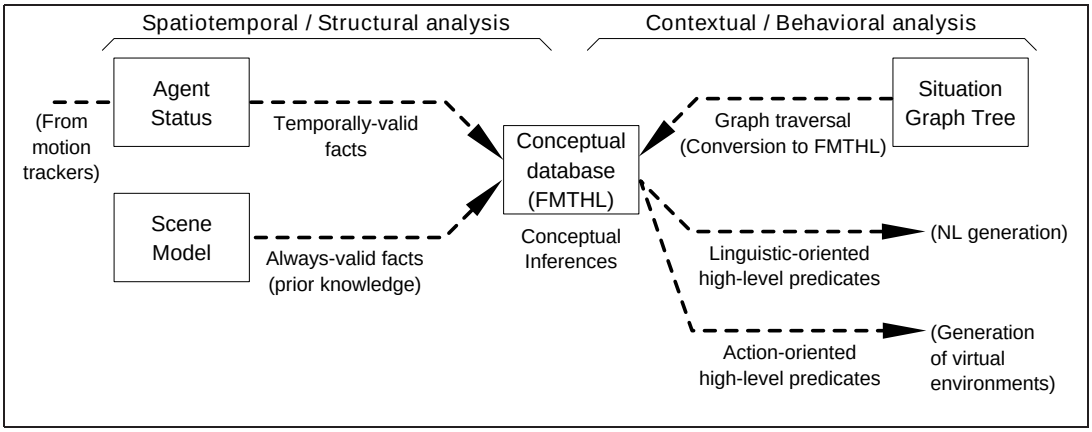


Fig. 4. The FMTHL inference engine manages (i) the conceptualization step for the motion data, (ii) the assertion of new facts given a series of atomic predicates, and (iii) the application of situational models from the BIL.

```
MaxDif < 30 ).
```

6 Behavior Interpretation

An independent stage is implemented for achieving 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, i.e. the contextualization and interpretation proposed in Section 4.

We introduce the concept of *High-Level Semantic Predicates* (HLSP) as those which express semantic relations among entities, at a higher level than metric-temporal relations. They result from applying situational models over STP. These new constraints embed restrictions based upon *contextualization*, *integration*, and *interpretation* tasks. Hence, the set of HLSP reaches the highest

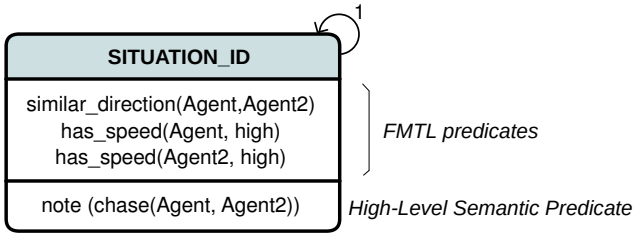


Fig. 5. Situation scheme from a SGT. In this example, a set of conditions in form of STP enable the generation of a HLSP, encoded in form of reaction predicates.

account of semantics, in the cognitive sense that each one of them implies a perceived situation or behavior which is meaningful and remarkable by itself in the selected domain. HLSP have been chosen as central elements in the semantic environment of the CVS, for them being language-independent and suitable for a neutral framework between vision and linguistics.

The tool which has been chosen to enable behavior modeling of the HLSP is the Situation Graph Tree (SGT), see [2,10]. The 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. 5. 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. 6, 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 situations. A part of a basic SGT is shown in Fig. 7, which illustrates a model to identify situations such as an abandoned object or a theft.

As previously shown in Fig. 4, 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 HLSP predicates are indexed with the temporal interval in which they have been the persistent output of the situational analysis stage. As a result, 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 interre-

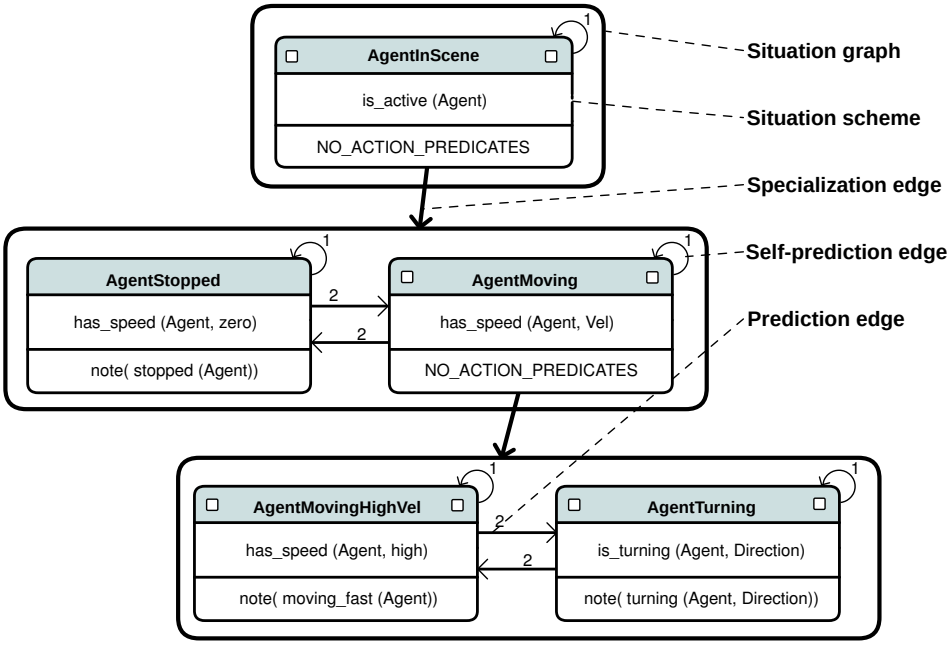


Fig. 6. A naive example of a SGT, depicting its main components. Specialization edges allow particularizing a general situation scheme with one of the situations within its child situation graph, in the case that more information is available. Prediction edges indicate the situations available from the current state for the following time-step; in particular, self-prediction edges hold a persistent state.

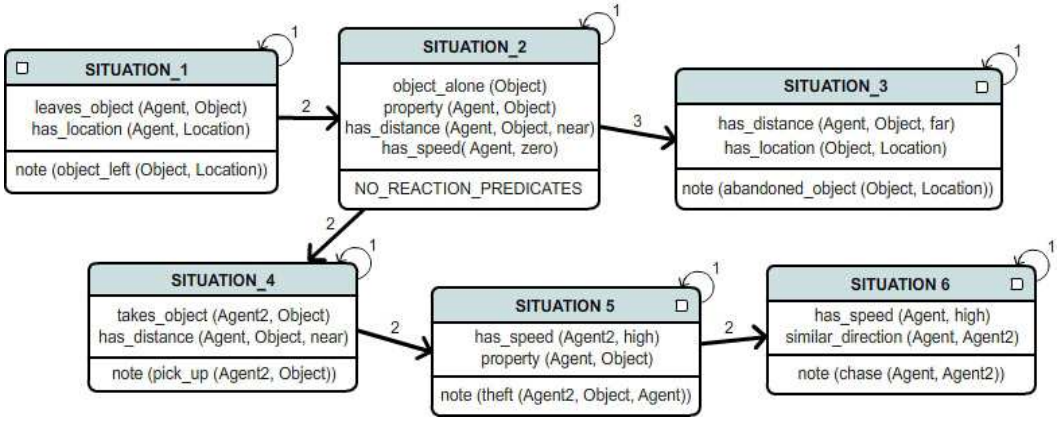


Fig. 7. This situation graph is evaluated when the system detects that an object has been left by the pedestrian who owns it. The set of conditions are FMTHL predicates, the reaction predicate is a `note` command which generates a high-level semantic tag.

lating 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 HLSP 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, a 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 HLSP 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 8.

7 User Interaction

A fundamental objective of cognitive systems is to achieve effective human-machine interaction, in order to enhance human capability and productivity across a large collection of endeavors related to a definite domain. Toward this end, several interfaces of communication are required, which make knowledge accessible to external users and permit them to take control over the implicit operations at each level. Ontologies play a fundamental role, especially regarding NL understanding tasks, since they restrict the domain of validity of the

linguistic queries and organize the knowledge to achieve an effective process of information retrieval.

In the case of the CVS discussed here, the multilingual interface of communication with a final user has been conceived to support two types of human-machine interactions:

(1) Generation of textual descriptions in natural language, in a multilingual environment.
(2) Understanding of user queries written in natural language, also for multiple languages.

NL generation has been seen as a process of choice, whereas NL understanding is best qualified as one of management of hypothesis towards reaching the most probable interpretation of a linguistic input [22]. A detailed description of the multilingual NL generation module for the system is presented in [9].

Regarding the NL understanding module, it has been considered convenient for it to enable communication with an external user at different levels of interaction:

- Answering questions about previous or current developments detected in a scene, such as "*How many pedestrians are there in the crosswalk at the moment?*", or "*Please tell me about any potentially dangerous behaviors occurred within the last 3 hours*".
- Carrying out specific commands to control and guide the system. A primary interest for the CVS is to act over its set of active cameras, so that the regions of interest can be focused and best exploited. Accepting sentences like "*Zoom in on the last pedestrian*" is interesting for this purpose. However,

more general directives should be possible, such as "*Concentrate only on the people appearing by the right side*".

- The user also has the possibility to enhance or modify the employed knowledge base, e.g. by including new information about the scenario or renaming identifiers.

The use of ontological semantics, and in particular the application of DL reasoning tasks over the relational database, has been proven to be an efficient solution when dealing with large amounts of data. Specifically, *instance checking* capabilities permit to perform tasks such as checking out whether all defined concepts admit individuals (consistency), finding the most specific concept an individual is instance of (realization), or retrieving the list of all individuals which instantiate a particular concept (retrieval), among others [3].

The ontology reduces the domain of validity undertaken by the universe of user queries, and makes it possible to restrict them to a reduced space of situations. By asserting concepts from the knowledge base, the instantiated elements from the ontology are related to particular knowledge sources, such as the *factual database* and the *onomasticon* [20], see Fig. 8. The factual database makes the information asserted at the ABox permanently available, so that an awareness of the occurred facts over time can be accessed as needed. The onomasticon consists of a repository of names, used to relate the instances of concepts from the ontology to the most suitable identifiers or referring expressions, keeping in mind to maintain the coherence of the semantic discourse. The onomasticon directly points to elements of the factual database, for instance stating that `Agent4` was the individual from the `Agent` class (defined in the ontology) which participated in a `chase` occurrence, instantiated in the factual database from frames 1241 to 1250 of an outdoor scene video sequence.

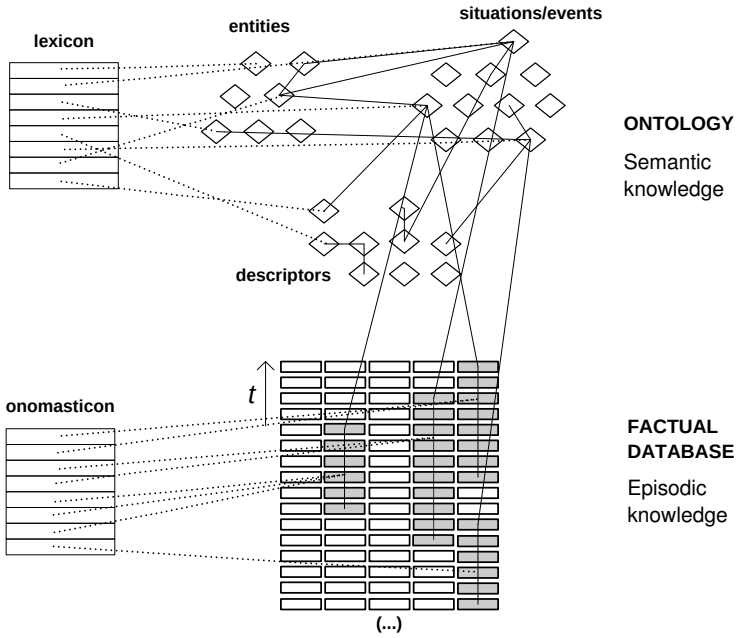


Fig. 8. When identifying a situation, the links among the different participants of the ontology, fundamentally entities, occurrences, and descriptors are inspected. Entries from the onomasticon and the lexicons are identified as afore known entities (such as names for particular locations or for agent categories). Instantiated situations are collected into the factual database for episodic awareness and remembering.

Fig. 9 shows a directed scheme of the process designed for the conversion from NL user queries to specific actions to carry out. First, the NL Understanding module parses the NL query basing uniquely on linguistic models. The a priori information from the scene, i.e. elements of the scenario and types of events or situations expected, is particularly provided by a multilingual lexicon. After that, the Dialog Manager module selects a suitable goal query by means of a pattern matching operation over the disambiguated tree structure resulting from parsing. The list of currently admissible goal queries for the system is presented in Table 4. The resulting goal query is forwarded to a Conceptual Reasoner, which performs a search over the factual database, aligns the information, and decides the action to perform according to the plan associated to

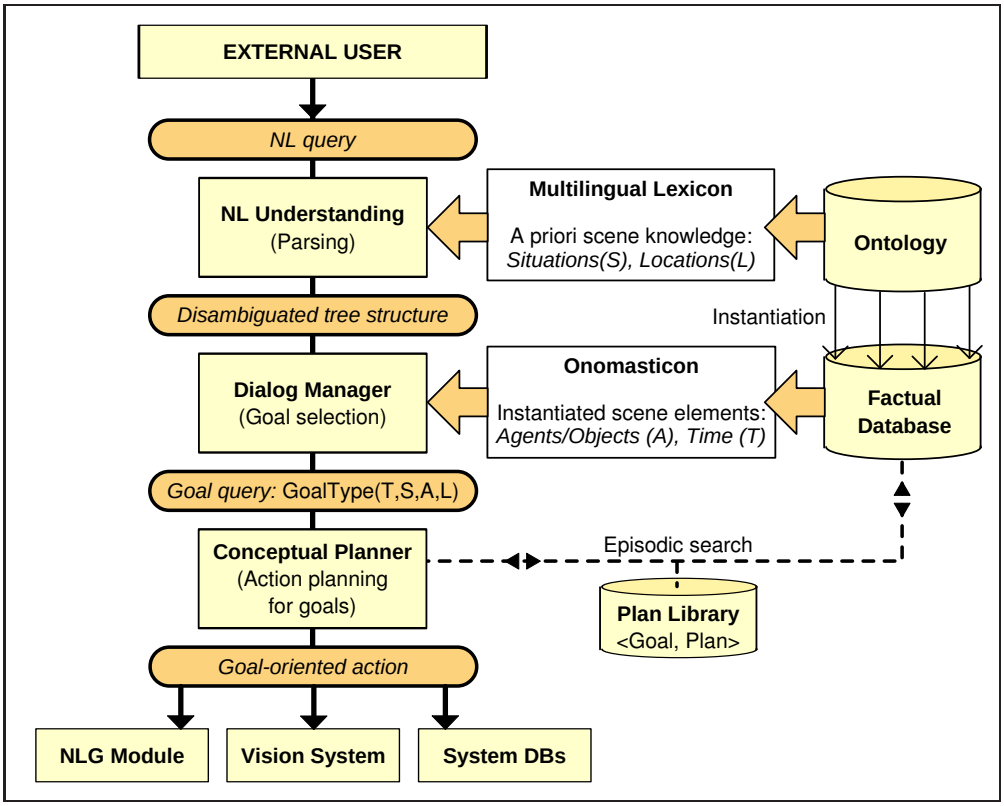


Fig. 9. Scheme of the process designed to transform NL user queries in goal-oriented actions.

each goal. The possible final actions include presenting information to the user by means of the NL generator module, forwarding specific actions to the active vision system, or modifying current information in the system databases.

8 Experimental results

Figs. 10 and 11 show current experimental results, in which a collection of HLSP have been successfully generated for sequences recorded in outdoor and indoor surveilled scenarios, respectively. [1] The collection of HLSP describe in-

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

Type of query	Admissible syntax	
QUERIES	list { (T), A }	*"How many thefts have been seen?"* list { S=bTheft }
	list { (T), S, (A), (L) }	
	assert { (T), S, (A), (L) }	*"Has there been any abandoned object?"* assert { S=bAbandonedObject }
	query { T?, (S), (A), (L) }	
	query { (T), S?, (A), (L) }	*"What is the last person doing in the table?"* query { S=?, A=Agent8, L=tables }
	query { (T), (S), A?, (L) }	
	query { (T), (S), (A), L? }	*"When was the machine kicked?"* query { T=?, S=aKick, L=VendingMachine }
COMMANDS	restrict { A }	
	restrict { L }	*"Concentrate on pedestrians appearing from the right part"* restrict { S=ceAppear, A=SinglePedestrian, L=right }
	restrict { (T), S, (A), (L) }	
	show { (T), (S) }	*"Show me persons appearing in the last 500 frames"* show { T=(668,1168), S=ceAppear, A=SinglePedestrian }
KNOWLEDGE BASE MODIFICATION	add { T, S, A, L }	
	remove { (T), (S), (A), (L) }	*"An object was abandoned on frame 1300"* add { T=(1300,1300), S=bAbandonedObject }
	rename { A }	

Table 4

List of admissible goal queries. In the current implementation, a goal query reduces the candidates in a semantic frame to a four-tuple consistent of Time (T), Situation (S), Agent/Object (A), and Location (L). Some examples for goal interpretations of NL input queries are shown, too.

teractions 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.

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 FMTHL 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 HLSP 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 75 possibly interesting NL questions or commands to formulate have been proposed by a group of 15 persons. 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*

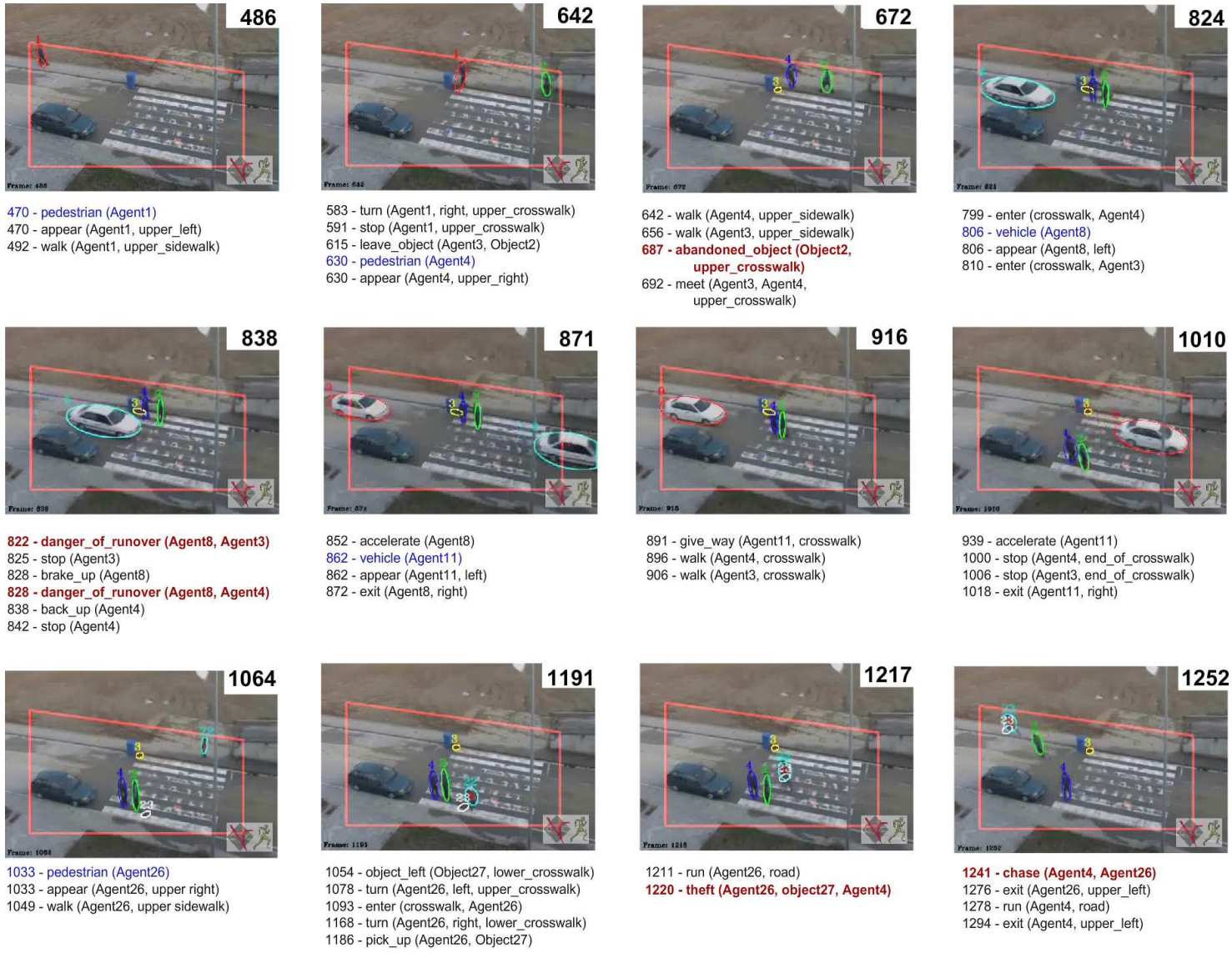


Fig. 10. Set of semantic annotations produced for the outdoor scene, which have been automatically generated for the fragment of recording comprised between frames 450 and 1301.

Fig. 11. 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.

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 FMTHL 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 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 pro-

vided 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 [2] inference engine for fuzzy metric-temporal horn logic and the SGTEditor [3] graphical editor for SGTs. On the other hand, the implementation of the ontology and the query system have been developed using the Protégé [4] ontology editor and the Jena [5] Semantic Web Framework.

9 Conclusions

We have presented the high-level architecture of a cognitive vision system for surveillance purposes, which intends to extract plausible interpretations of occurrences appearing in video sequences from a definite domain. The CVS presented is designed in structured levels of reasoning, that allows extracting interpretations at different semantic levels. In addition, a relational database management system has been described in a first step, which has been considered from an ontological semantics perspective, in order to offer broad practical coverage for complex situations requiring to process large amounts of data.

The proposed taxonomical building of the events in a domain implies different levels of interpretation, ranging from basic actions and events (e.g. walk, run, turn) to contextualized events and a more scenario-specific interpretation of behaviors (e.g. meeting, giving way, chasing). The basic levels for the status

[2] http://cogvisys.iaks.uni-karlsruhe.de/Vid-Text/f_limette/index.html
[3] http://cogvisys.iaks.uni-karlsruhe.de/Vid-Text/sgt_editor/index.html
[4] http://protege.stanford.edu/
[5] http://jena.sourceforge.net/

of agents are defined using human motion models, thus being applicable in a general way. The compilation of new behaviors for a new scene is especially focused on increasing or modifying the other levels, which are more dependant on the particular working scenario.

Modeling an ontology permits to reduce the complexity associated to the multilingual dimension of the system. It also allows to clarify and simplify the design and implementation of components to bridge the semantic gap. Ontologies are particularly useful in the application field of NL understanding, since they make easier the categorization of a discourse and play a great role in disambiguation. Another important benefit is to restrict the domain of acceptance for the different forms of semantic representation, since the constraints applied to the terminology fix the validity of the situations to detect. This way, mechanisms for prediction based on restrained behavioral models can be developed.

As stated in [20], changes in the topology of the ontological hierarchy and in the distribution of knowledge do not hold special significance. What has been realized as much more crucial has been to focus on coverage, in order to find the most suitable grain size of the possible semantic representations with relation to the needs of the concrete application, e.g. how particular have to be the interpretations in order to detect a given set of occurrences. The main idea is to model those behaviors requiring a more subjective interpretation in a way such as they are not wrongly extended to general situations, and also making easy deducible situations not require excessively detailed information. 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.

By increasing the complexity of the scenes in terms of number of cameras and number of agents, we actually increase the difficulties to the tracking systems. If the vision levels concerning tracking and calibration tasks are robust enough, the upper levels of the system should be able to deal with this kind of complexity, only by paying a higher computational time. Nevertheless, the behavior of crowds or large groups of agents has not been analyzed yet, and it has to be included as future work: the exponential growing of complexity for certain situations could easily complicate real-time performance at the semantic levels. In order to solve such difficulties, new approaches need to be investigated, firstly involving the exploitation of confidence levels resulting from the conceptual reasoning stage.

The addition of alternative domains, such as soccer, will be tackled in the future. In this regard, one of the most time consuming tasks will be deciding upon a final taxonomy of accepted situations, which is both based on the available tracking capabilities and the complex design of the situational models. Managing large group interactions and prioritizing the situations of interest are issues that will probably represent a great cost, too, once the functionality of the tracking system is granted. The rest of tasks follow the described approach.

One immediate application is related to the field of semantic indexation. The ontology of situations provides the space and validity of possible annotations for video sequences related to the domain. In the specific implementation, a SGT acts as an actual content classifier, which characterizes the temporal structure of video sequences from a semantic perspective. Thus, the resulting

predicates can be identified as high-level semantic indexes, which can facilitate further applications such as search engines and query-based retrieval of content.

The approach suggests enhancements for the automatization of many tasks, such as the automatic categorization of semantic regions in the scenario, or regarding discourse goal recognition at a pragmatic level, for example. Experimental results show that the bidirectional collaboration between ontological resources and reasoning stages at different representation levels provides expressive results. Nevertheless, one fundamental requirement is to establish a pondered trade-off between the use of reasoning capabilities, which enable powerful but rigid interpretations, and the flexibility of the relational databases, in which an increase of complexity notably diminishes the effectiveness of the cognitive system.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

[1] Abiteboul, S., Hull, R., and Vianu, V., Foundations of Databases. 1995, Addison Wesley Publi. Co., London.

[2] Arens, M. and Nagel, H.–H., Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences. 2003, Proc. of the KI 2003, 2821(1):149–163. Springer–Verlag, Berlin, Heidelberg, New York.

[3] Baader, F., Calvanese, D., McGuiness, D., Nardi, D., and Patel-Schneider, P. (eds.), The Description Logic Handbook. 2003, Cambridge University Press, Cambridge.

[4] Baiget, P., Fernández, C., Roca, X., and Gonzàlez, J., Automatic Learning of Conceptual Knowledge for the Interpretation of Human Behavior in Video Sequences. 2007, Proc. of the 3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007), 4477(1):507–514, Springer LNCS, Girona, Spain.

[5] Baiget, P., Soto, J., Roca, X., and Gonzàlez, J., Automatic Generation of Computer Animated Sequences Based on Human Behavior Modeling. 2007, Proc. of the 10th 3IA International Conference in Computer Graphics and AI (3IA 2007).

[6] Baker, C. F., Fillmore, C. J., and Lowe, John B., The Berkeley FrameNet project. 1998, in Proc. of the COLING-ACL, Montreal, Canada.

[7] Buxton, H., Learning and Understanding Dynamic Scene Activity: a Review. 2003, Image and Vision Computing, 21(1):125–136. Elsevier Science.

[8] Crowley, J.L., Situated Observation of Human Activity. 2005, 1st Workshop on Computer Vision for Interactive and Intelligent Environment (CVIIE 2005), pp.97–108. IEEE Computer Society, Lexington, Kentucky, USA.

[9] Fernández, C., Baiget, P., Roca, X., & Gonzàlez, J., Natural Language Descriptions of Human Behavior from Video Sequences. 2007, 30th Annual German Conference on AI (KI 2007), Osnabrück, Germany. Springer LNCS.

[10] Gonzàlez, J., Human Sequence Evaluation: The Key-Frame Approach. PhD Thesis. ISBN 84–933652–2–X. 2004, Universitat Autònoma de Barcelona, Barcelona, Spain.

[11] Granlund, G., Cognitive Vision Systems. Organization of Architectures for Cognitive Vision Systems. 2006, Nagel, H.–H. and Christensen, H.I. (eds.), pp. 37–55. Springer Verlag, Heidelberg.

[12] Haag, M., Theilmann, W., Schäfer, K., & Nagel, H.–H., Integration of Image Sequence Evaluation and Fuzzy Metric Temporal Logic Programming. 1997, Proc. of the 21st Annual German Conference on AI, pp. 301–312. Springer-Verlag, London, UK.

[13] Huerta, I., Rowe, D., Mozerov, M., & Gonzàlez, J., Improving Background Subtraction based on a Casuistry of Colour-Motion Segmentation Problems. 2007, Proc. of the 3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007), Girona, Spain. Springer LNCS.

[14] Kojima, A. and Tamura, T. & Fukunaga, K., Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. 2002, International Journal of Computer Vision, 50(2):171–184.

[15] Kompatsiaris, I., Avrithis, Y., Hobson, P., & Strinzis, MG, Integrating Knowledge, Semantics and Content for User-Centred Intelligent Media Services: the aceMedia Project. 2004, Proc. of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS04), Lisboa, Portugal, pp. 21–23.

[16] Ma, M. & Mc Kevitt, P., Visual semantics and ontology of eventive verbs. 2004, Proc. of the 1st International Joint Conference on Natural Language Processing, 1(1):278–285.

[17] Marburger, H. and Neumann, B., & Novak, H.J., Natural Language Dialogue about Moving Objects in an Automatically Analyzed Traffic Scene. 1981, Proc.

IJCAI-81, Vancouver.

[18] Nagel, H.–H. and Gerber, R., Representation of Occurrences for Road Vehicle Traffic 2008, AI-Magazine, 172(4-5):351–391.

[19] Nagel, H.–H., Steps toward a Cognitive Vision System. 2004, AI-Magazine, 25(2):31–50.

[20] Nirenburg, S. & Raskin, V., Ontological Semantics. 2004, MIT Press.

[21] Park, S. & Aggarwal, J.K., Event Semantics in Two-person Interactions. Proc. of the Pattern Recognition, 17th International Conference on (ICPR'04). 2004, 4(1):227-230. IEEE Computer Society Washington, DC, USA.

[22] Reiter, E. & Dale, R., Building Natural Language Generation Systems. 2000, Cambridge University Press, Cambridge/UK.

[23] Rowe, D., Rius, I., Gonzàlez, J., & Villanueva, J.J., Improving Tracking by Handling Occlusions. 2005, 3rd ICAPR, UK, 2(1):384-393. Springer LNCS.

[24] Schäfer, K., Brzoska, C., "F-Limette" Fuzzy Logic Programming Integrating Metric Temporal Extensions. 1996, Journal of Symbolic Computation, 22(5-6):725–727. Academic Press, Inc. Duluth, MN, USA.

[25] Talmy, L., Toward a Cognitive Semantics - Vol. 1: Concept Structuring Systems. 2000, Bradford Book.

[26] Vincze, M., Ponweiser, W. & Zillich, M., Contextual Coordination in a Cognitive Vision System for Symbolic Activity Interpretation. 2006, Proc. of the 4th IEEE International Conference on Computer Vision Systems, 1(1):12–12. IEEE Computer Society Washington, DC, USA.

[27] Wilson, R.A. & Keil, F.C. (Editors), The MIT Encyclopedia of the Cognitive Sciences. 2001, Bradford Book.