High-level Integration for Cognitive Vision Surveillance

Carles Fernández*, Pau Baiget*, Xavier Roca*, Jordi Gonzàlez+

* *Computer Vision Centre, Edifici O. Campus UAB, 08193, Bellaterra, Spain*
+ *Institut de Robòtica i Informàtica Ind. UPC, 08028, Barcelona, Spain*
E-mail: perno@cvc.uab.es

Abstract The increasing interest in Cognitive Vision Systems (CVS) motivates the apparition of ad-hoc stages designed for the integration of multiple kinds of knowledge. This paper proposes a novel ontology to restrict and integrate high-level semantics for Human Sequence Evaluation (HSE), which targets multilingual capabilities and multipurpose end-user interfaces. The main contributions of this paper are the conception of a neutral semantic layer, which allows to link vision and linguistic domains; and the use of *situations* instead of verbs as basic elements for an ontological categorization of occurrences. In our approach, the domain has been restricted to outdoor surveilled scenarios, involving interactions among pedestrians, static objects, and vehicular traffic.

1 Introduction

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.

The proposed Cognitive Vision System is based upon the conception of *Human Sequence Evaluation* (HSE), in which the interpretation of human behaviors in image sequences is performed by a modular architecture for user-oriented applications [2]. In such a framework, it is essential to develop proper criteria for high-level knowledge sharing and validation. Due to the broad spectrum of semantic representations, it is necessary to find mechanisms that clarify the structure of knowledge in given domains, for integration purposes. Towards this end, ontologies have been widely accepted as convenient tools.

This contribution addresses the use of ontologies as an integrative framework for knowledge representation, within a HSE system with multiple user interfaces and multilingual capabilities. The goal is to automatically extract behavioral descriptions from image sequences in restricted domains, in this case urban outdoor surveillance environments. We also discuss criteria to model the semantic background of such ontologies, which link the different representations at high-level stages. The integration of cognitive capabilities for the aimed system has been thought to be implemented in a modular, collaborative distribution, in which the entailed tasks range from low-level, vision-related approaches, to high-level, conceptual and linguistic implementations. This paper discusses the various representation formalisms of the last stages, and solutions towards their collaboration. The proposed high-level architecture is shown in Fig. 1.

2 Representation Formalisms in HSE

Several formalisms are employed by a HSE system in order to represent semantic knowledge, which are conditioned to the application domain they address. Table 1 contains a summary of some remarkable features for the different semantic representation formalisms described.

2.1 Spatiotemporal Predicates (STP)

These predicates rely on Fuzzy Metric-Temporal Horn Logic (FMTHL), which facilitates a schematic

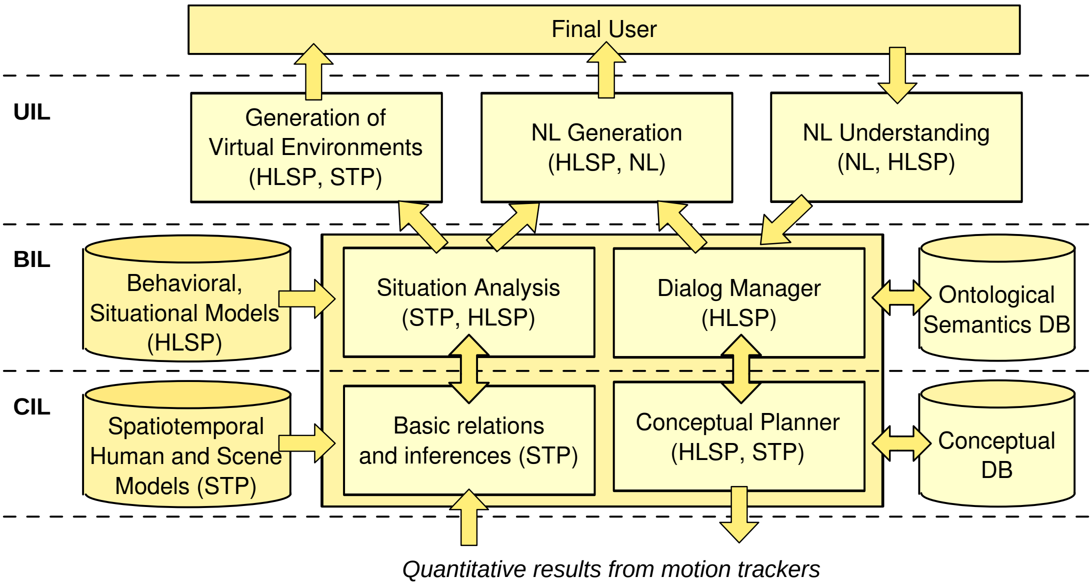

Final User
UIL
Generation of Virtual Environments (HLSP, STP)
NL Generation (HLSP, NL)
NL Understanding (NL, HLSP)
BIL
Behavioral, Situational Models (HLSP)
Situation Analysis (STP, HLSP)
Dialog Manager (HLSP)
Ontological Semantics DB
CIL
Spatiotemporal Human and Scene Models (STP)
Basic relations and inferences (STP)
Conceptual Planner (HLSP, STP)
Conceptual DB
Quantitative results from motion trackers

Figure 1: Proposed high-level architecture for the HSE system. The knowledge representation formalism used by each module is enclosed in parenthesis. The different stages consist of Conceptual Primitives (CPL), Behavior Interpretation (BIL), and User Interaction (UIL) levels.

representation of conceptual knowledge which is time-delimited and incorporates uncertainty [6]. We use it to represent and reason about spatiotemporal developments, by assigning fuzzy degrees-of-validity to quantitative values generated by the motion trackers.

In the current implementation, FMTL is manipulated by the inference engine *F-Limette* [6] to represent and reason about spatiotemporal developments. Uncertainty is treated by assigning fuzzy degrees-of-validity to the quantitative values generated by the motion trackers [5]. Next example shows a metric-temporal modeling for the inference of a new FMTL predicate upon the quantitative values for the orientations of two agents.

```
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
).
```

2.2 High-Level Semantic Predicates (HLSP)

High-level Semantic Predicates are thought to 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 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.

Our implementation for the generation of these high-level predicates is based on Situation Graph Trees (SGTs), see [1]. The nodes of these graphs are schemes which embed the contextual state of an agent at a discrete point of time, by relating a set of necessary FMTL facts to the situation. When the entire set of facts defined is asserted, a new interpretation for the scene is generated in form of a HLSP. SGTs are traversed at every time-step, and therefore the produced interpretations in HLSP are subjected to temporal validity.

2.3 Linguistic Predicates (LP)

These predicates represent linguistic-oriented knowledge. They are incorporated using Discourse Representation Theory (DRT) [3]). They are used for NL generation and understanding. Each LP requires distinct thematic arguments depending on the language

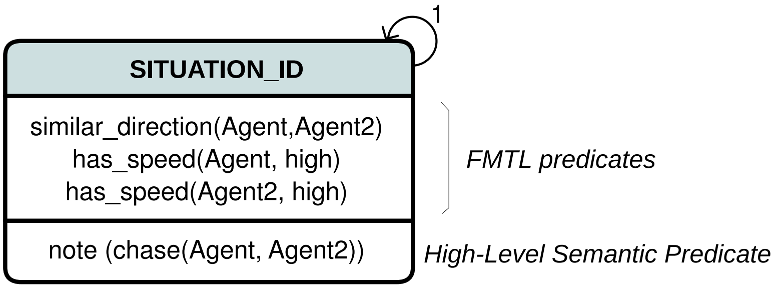

1
SITUATION_ID
similar_direction(Agent,Agent2)
has_speed(Agent, high)
has_speed(Agent2, high)
FMTL predicates
note (chase(Agent, Agent2))
High-Level Semantic Predicate

Figure 2: Situation scheme from a SGT.

and situation. LP in different languages describing a single situation are related to a single HLSP.

A Discourse Representation Structure is constituted by a set of referents and a universe of conditions. In our case, the referents are chosen from the set of instantiable entities which will be defined by the ontology, and the conditions are conformed by a subset of predicates linguistically oriented towards a particular language. Each of these LP requires distinct thematic arguments depending on the language and situation. An example of DRS including LP in English is shown next for the sentence *"A theft was detected"*. [1].

$$e_1 : \langle \{x, n, t_1, e_1\}, \{theft(x), e_1 : detect(x), t_1 < n, e_1 \subseteq t_1\} \rangle$$

We focus on HLSP for building the ontology, for them being language-independent and suitable for a neutral framework between vision and linguistics. Fig. 3 shows a collection of HLSP which have been successfully generated for a sequence recorded in an outdoor surveilled scenario, involving pedestrians, pickable objects, and vehicular traffic. The collection of HLSP describe interactions among these entities.

3 Ontologies for integration of knowledge

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 [4], from which we remark Case Grammar, Lexical Conceptual Structures, Thematic Proto-Roles, WordNet, Aspectual Classes, and Verb Classes. As an extension, our approach relates each situation from the ontology with a set of required entities, which are classified depending on the thematic role they develop. The main advantage of this approach in an independency of the particularities of verbs to a concrete natural language, thus facilitating addition of multiple languages.

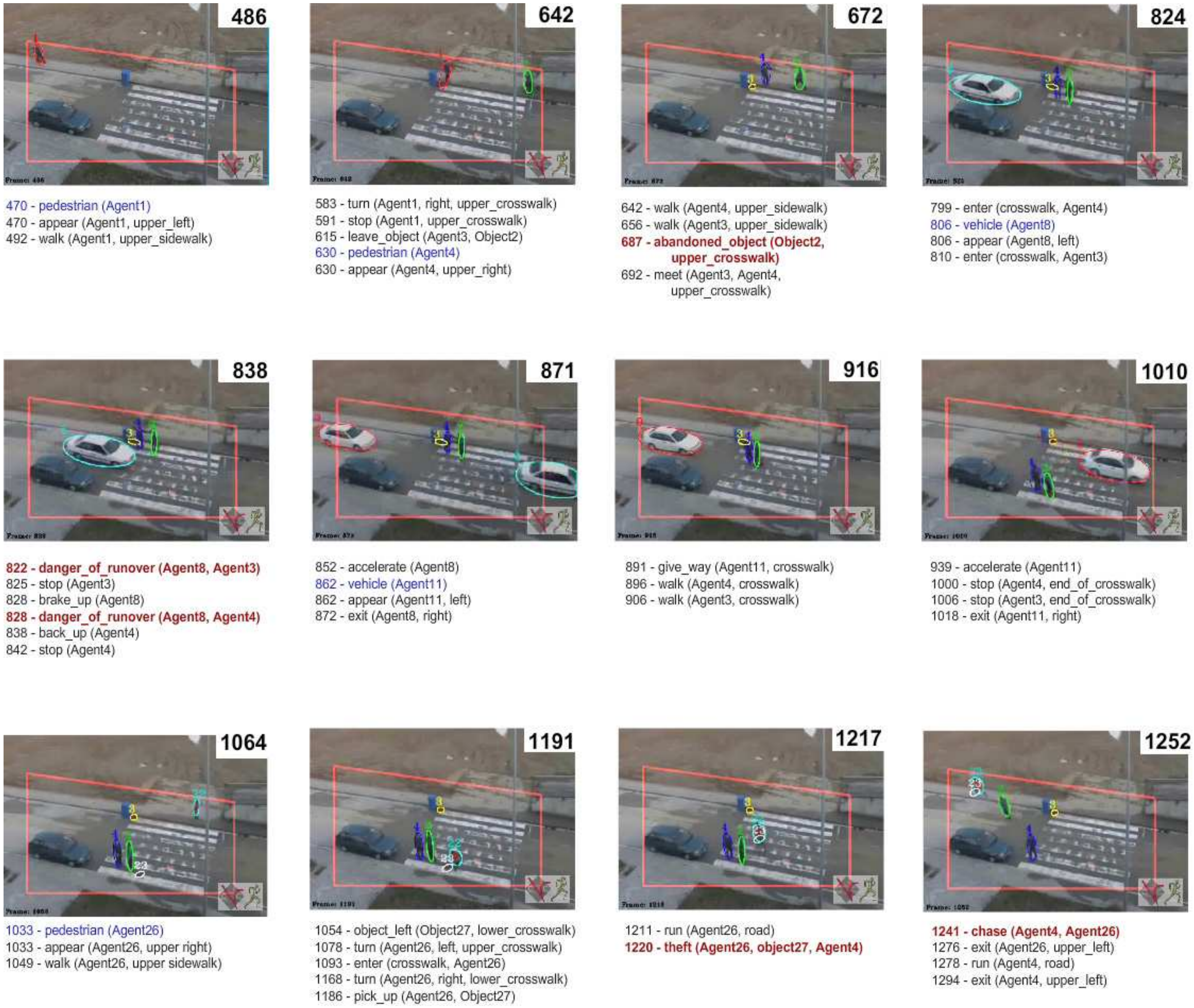
Another taxonomy defines a set of semantic entities in the domain. The chosen list includes *agents* as those which can spontaneously act to change a situation, here pedestrians and vehicles; *objects* as static elements of the scene; *locations*; and also a set of abstract *descriptors* which permit to add fuzzy modifiers to the conditions related to the entities. Other roles such as experiencer, goal, location, or instrument are easily enclosed in the selected categories.

3.1 Ontological Categorization of Situations

The main target for the proposed ontology is to enumerate and correlate the instantiable situations which are detectable in the selected domain, using a proper cognitive-based semantic representation. Now that the possible semantic participants have been established and organized, the set of situations can be classified.

Talmy organizes conceptual material in a cognitive manner by analyzing what he considers most

[1] The condition between temporal referents $t_1 < n$ characterizes the past tense for the NL generation.



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

	STP	HLSP	LP
Type of semantics	Metric-temporal (basic relations)	Thematic roles (inferential role semantics)	Linguistic-oriented (NL semantics)
Models implied	Scene models, human motion models	Behavioral models (contextual and intentional)	Linguistic models (Syntax, morphology, alignment, etc.)
Benefits	Allows inference of higher-level predicates upon asserted facts	Linguistic-oriented, highest level of interpretation	Facilitates to convert between logic and NL
Limitations	Limited to metric-temporal reasoning	Domain-dependent and target-oriented	Language-dependent

Table 1: Table of semantic representations in HSE.

crucial parameters in conception: space/time, motion/location, causation/force interaction, and attention/viewpoint [7]. For him, semantic understanding involves the combination of these domains into an integrated whole. Our classification of situations agrees with these structuring domains: We organize semantics in a linear fashion, ranging from objective knowledge in vision processes (low-level) to uncertain, subjective knowledge based on attentional factors (high-level). It is structured as follows, see Table 2:

- The *Status* class contains metric-temporal knowledge, based on the information provided by the considered trackers: body, agent, and face. Its elements represent spatial configurations and analysis of agent trajectories.
- The *ContextualizedEvent* class involves semantics at a higher level, now considering interactions among semantic entities. This knowledge emerges after contextualizing different sources of information, what 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.

Each of the described behaviors requires certain arguments, characterized by the mentioned entities. For instance, a *DangerOfRunover* situation involves at least two Agents, a Vehicle and a Pedestrian, and a *Theft* situation involves a minimum of two Pedestrians and an object of type *PickableObject*.

4 Conclusions and Future Work

An ontology has been designed to account and organize the universe of situations to be handled by a CVS for surveillance purposes. These situations are represented by HLSP, which hold a high level of semantics and are language-independent. The resulting ontology builds on a neutral framework between vision and linguistics. The proposed modeling is particularly useful for multilingual NL interfaces, making easier tasks of discourse categorization and disambiguation. It also restricts the domain of acceptance for semantic formalisms, facilitating prediction. One direct application is related to semantic indexation: The set of HLSP can be seen as the universe of high-level indexes in a domain, which facilitate further applications such as search engines and query-based retrieval of content. Several issues must be covered in next steps: proper communication between the semantic layer and the NL interface requires to relate the proposed ontology of situations to a linguistic-oriented one. In addition, the domains of application have to be enlarged.

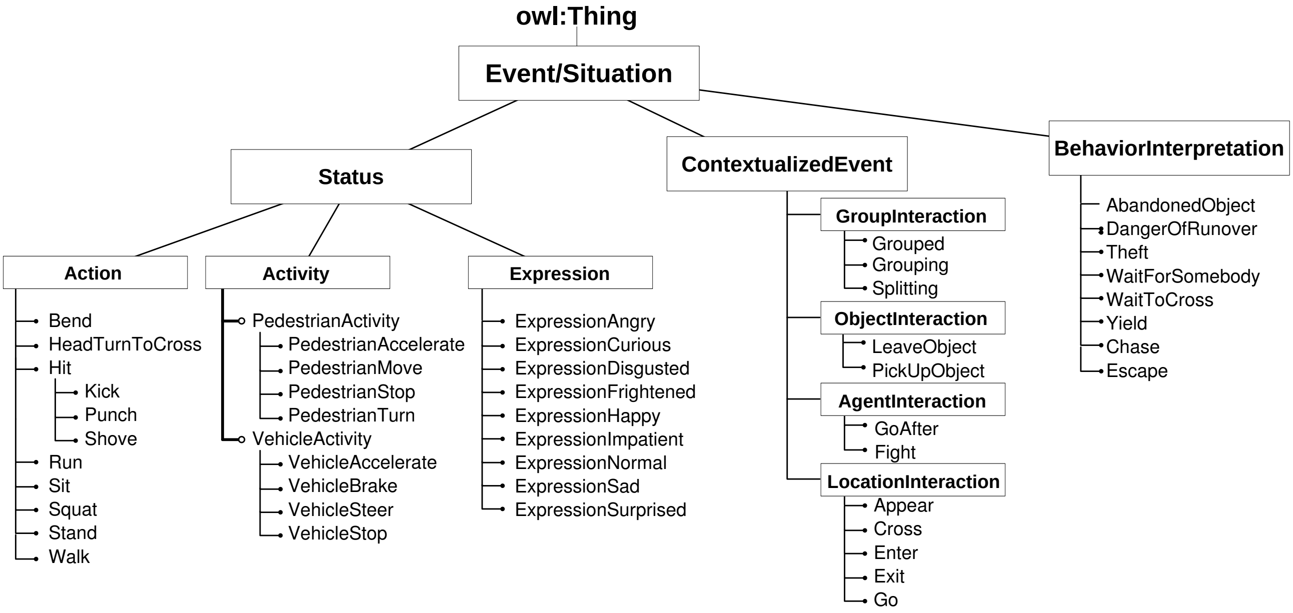


Table 2: Central part of the ontology: the taxonomy for a classification of situations.

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 DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

[1] M. Arens and H.-H. Nagel. Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences. volume 2821, pages 149–163. Springer-Verlag Berlin, Heidelberg, New York, 2003.

[2] J. Gonzàlez. *Human Sequence Evaluation: The Key-Frame Approach*. PhD thesis, Universitat Autonoma de Barcelona, Barcelona, Spain, 2004.

[3] H. Kamp and U. Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht Boston London, 1993.

[4] M. Ma and P. Mc Kevitt. Visual semantics and ontology of eventive verbs. *Proc. of the 1st International Joint Conference on NL Processing*, pages 278–285, 2004.

[5] D. Rowe, I. Huerta, J. Gonzlez, and J.J. Villanueva. Robust Multiple-People Tracking Using Colour-Based Particle Filters. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007)*. Springer LNCS, 2007.

[6] K. Schäfer and C. Brzoska. F-Limette Fuzzy Logic Programming Integrating Metric Temporal Extensions. *Journal of Symbolic Computation*, 22(5-6):725–727, 1996.

[7] L. Talmy. *Toward a Cognitive Semantics - Volume 1*. Bradford Book, 2000.