Natural Language Descriptions of Human Behavior from Video Sequences

Anonymous KI2007 submission

Abstract. This contribution addresses the generation of natural language textual descriptions in Catalan, English, and Spanish for real-time evaluation of human behavior in video sequences. The problem is tackled by converting geometrical information extracted from videos of the scenario into predicates in fuzzy logic formalism, which facilitates the internal representations of the conceptual data and allows the temporal analysis of situations by means of Situation Graph Trees. The results of the analysis are stored in Discourse Representation Structures, which facilitate a subsequent generation of natural language text. This set of tools has been proved to be perfectly suitable for the specified purpose.

1 Introduction

The introduction of Natural Language (NL) interfaces into vision systems is becoming popular, especially for surveillance systems. Methods for describing human activities from video images have been reported by Kojima and Tamura [6], and automatic visual surveillance systems for traffic applications have been studied by Nagel [7] and Buxton and Gong [2] among others. In a visual surveillance system, human behavior is represented by scenarios, i.e. predefined sequences of events. The scenario is evaluated and automatically translated into text by analyzing the contents of the image over time, and deciding on the most suitable predefined event that applies in each case.

Natural language text generation for the evaluation of human behavior in video sequences builds upon three disciplines, namely computer vision, knowledge representation, and computational linguistics. Thus, the overall architecture of the proposed system consists of three subsystems, see Fig. 1; a Vision Subsystem (VS), which provides the geometric information extracted from a video sequence by means of detection and tracking processes, a Conceptual Subsystem (CS), which infers the behavior of agents from the conceptual primitives based on the geometric information extracted by the VS, and a Natural Language Subsystem (NS), which in principle comprises the natural language text generation, but also becomes a good stage for providing a complete interface of communication with a final user [7].

Due to space limitations, the extraction of visual information is not treated here. Details can be found, for example, in [9]. We proceed on the basis that structural information consisting of geometrical values are available over time.

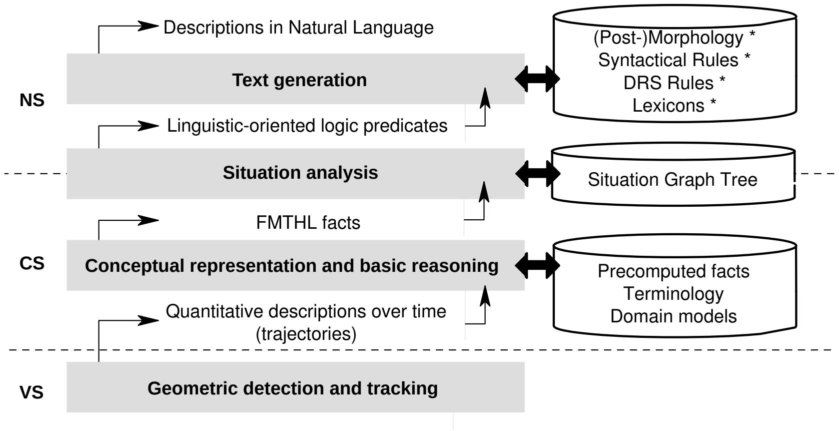
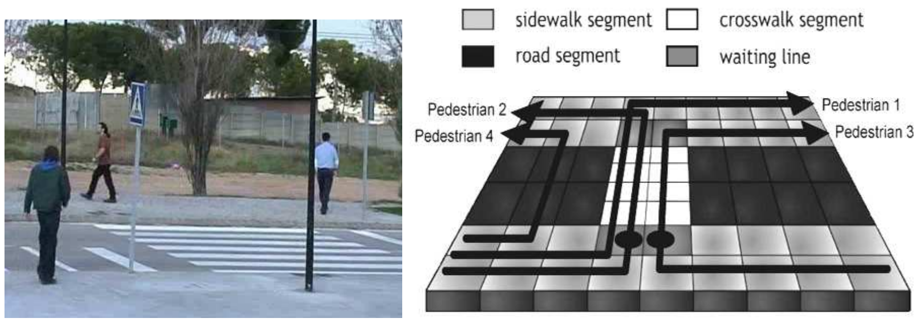


Fig. 1. General schema of the stages and interfaces related to the current text generation system. The left acronyms represent different sub-systems, the boxes describe the main processes that produce changes in data representations, and the right components specify some of the external tools required by the processes. An asterisk remarks that a resource is language-dependent.

The obtention of knowledge derived from visual acquisition implies a necessary process of abstraction. In order to understand the quantitative results from vision, it becomes fundamental to reduce the information to a small number of basic statements, capable of detecting and relating facts by means of qualitative derivations from what has been ‘seen’. The conversion of observed geometrical values over time into predicates in a fuzzy logic formalism allows to reach an intermediate state, the conceptual representation layer, which facilitates schematic representations of the scenarios [1] and, in addition, enables characterizations of uncertain and time-dependent data extracted from image sequences. Next, a classification can be performed by clustering these resulting facts into preconceived patterns of situations. Such an inference system produces not only an interpretation for the behavior of an agent, but also reasons for its possible reactions and predictions for its future actions [4].

Discourse Representation Theory seems to be of particular interest for the conversion from conceptual to linguistic knowledge, since it discusses algorithms for the translation of coherent natural language textual descriptions into computer-internal representations by means of logical predicates [5]. The reverse step is also possible, so that the results of the conceptual analysis are stored into semantic containers, the so-called Discourse Representation Structures (DRS), which facilitate the construction of syntactical structures containing some given semantic information. A final surface realization stage over these preliminary sentences embeds the morphological and orthographical features needed for obtaining final natural language textual descriptions.



(a) (b)

Fig. 2. Original pedestrian crosswalk scene (a) and groundplane schematic map of the main regions considered in this scene (b). Pedestrian trajectories have been included in the scheme. Black circles represent a stop on the waiting line.

2 Evaluation of Human Sequences

The chosen scenario for evaluation of basic human behaviors has been a crosswalk, see Fig. 2. On it, a certain number of pedestrians, each one with a different behavior, start from one of the sidewalks and cross the road to get to the other side. At first, the presence of traffic vehicles has been omitted.

2.1 The Conceptualization Step

The structural knowledge acquired by the VS needs to be abstracted and converted into logic knowledge in order to facilitate further manipulations and reasonings. To do so, trajectories and other types of estimated spatiotemporal information are associated to basic concepts identifying recognizable simple actions, which can be described by using elementary verb-phrases (e.g., 'approaching_to_location', 'turning', 'has_speed'). These *conceptual predicates* are not yet proper linguistic expressions, but system-internal representations resulting from classification and abstraction processes.

Fuzzy Metric Temporal 'Horn' Logic (FMTHL) has been conceived as a suitable mechanism for dealing with uncertain, time-dependent information [10]. This formalism allows to represent knowledge explicitly and hierarchically, not coded into conditional probabilities, and enables to manage data requiring both *temporal* and *fuzzy* properties [4]. Therefore, after an abstraction process is carried out, the reasoning system system is conferred a capability for representing uncertain qualitative descriptions inferred from the quantitative, geometrical values, e.g. postures, velocity, or positions for the agents, that have been acquired by analyzing agent trajectories within a predefined scene model. The logic productions evolve over time as the received data does, so this conceptual knowledge is also time-delimited, and thus the development of events can be comprehended and even anticipated.

The qualitative knowledge extracted from quantitative results is encoded using these fuzzy membership functions, so that the generated predicates are related to conceptual 'facts' for each time-step. For example, a collection of positions over time allows to derive fuzzy predicates such as 'has_speed(zero)', 'has_speed(small)', or 'has_speed(very_high)', depending on the displacements of the agent detected between consecutive points of time.

2.2 Agent Trajectories

Trajectory files are ordered collections of observed values over time for a certain agent, which are obtained as a result of the tracking processes for the agents [9]. From the evolution of the states of the agent, a certain *behavior*, i.e. a sequence of situated actions, will be assumed. Four agent trajectories have been obtained, which consist of a set of FMTHL logical predicates of type `has_status`. These predicates comprise the required knowledge for the human behavior analysis in the following scheme or *status vector* for the agent:

```
has_status(Agent, X0, Y0, Theta, Vel).
```

As can be seen, the `has_status` predicates for the interpretation of human actions contain five fields so far, all of them being identifiers to entities and objects detected during the tracking process, or otherwise concrete geometrical values for spatiotemporal variables. The `Agent` field gives information about the name given to the agent. The rest of the fields give quantitative values to the geometrical variables needed: 2-D spatial position in the ground plane (`X0, Y0`), angle of direction (`Theta`), and instant velocity (`Vel`). The `Vel` field provides the necessary information for determining the action being performed by the agent (i.e. *standing*, *walking*, *running*).

2.3 Scene Modeling

The scenario in which pedestrians perform their actions has been included as an additional source of knowledge for the reasoning stage. The geometrical modeling of the location has been done first in a ground plane bidimensional approach, so a set of spatial descriptors are declared to distinguish the relevant topographic or interesting elements in the scene, see Fig. 2 (b). This source provides the spatial distribution taken into account for the given situation.

A second source of knowledge contains other logical statements that will confer semantic significance on the initial geometrical descriptors of the scene. The different regions can be enclosed into different categories (*sideway*, *road*, *crosswalk*) and can also be given different attributes (*walking zone*, *waiting line*, *exit*). This step is necessary for identifying significative regions, so the movements and interactions of the agents can be contextualized by means of valid identifiers. These geometrical considerations have been encoded using FMTHL predicates.

2.4 Situation Graph Tree

Situation Graph Trees (SGTs) are hierarchical trees used to model the knowledge required from human behavior analysis in a specific discourse domain [3]. The conceptual knowledge about a given actor for a given time step is contained in a so-called *situation scheme*, which constitute the basic components of a SGT. The knowledge included in these components is organized in two fields:

- First, a set of logic conditions describes the requirements that need to be accomplished to instantiate that situation. The assertion of these conditions is performed by evaluating the semantic predicates inferred from the agent status vectors obtained at the visual stage.

- After the conditions have been asserted, certain domain-specific reaction predicates are generated and forwarded for defined purposes. Only generation of natural language text will be considered here, so *linguistic-oriented* logic predicates will be generated.

A single SGT incorporates the complete knowledge about the behavior of agents in a discourse [1]. Every possible action to be detected has to be described in the SGT. Consequently, it is necessary to have accuracy to precisely identify the desired actions, but it is also important that it does not become excessively complex in order to avoid a high computational cost. On the other hand, the SGTs are transformed into logic programs of a FMTHL for automatic exploitation of these behavior schemes. A situation graph tree has been designed for the crosswalk scene, see Fig. 3.

Depending on the behavioral state, a new high-level predicate will be sent to the NS Subsystem, by means of a `note` method. The new predicates offer language-oriented structures, since their attribute scheme comprises fields related to ontological categories such as *Agent*, *Patient*, *Object* or *Event*. These predicates are the inputs for the NS Subsystem, which will be discussed next.

3 Linguistic Implementation

It is in the NS where the logical predicates are used to provide the representational formalism, making use of the practical applications of the Discourse Representation Theory. Inside the NS layer, there are several stages to cover:

1. Lexicalization
2. Discourse Representation
3. Surface Realization

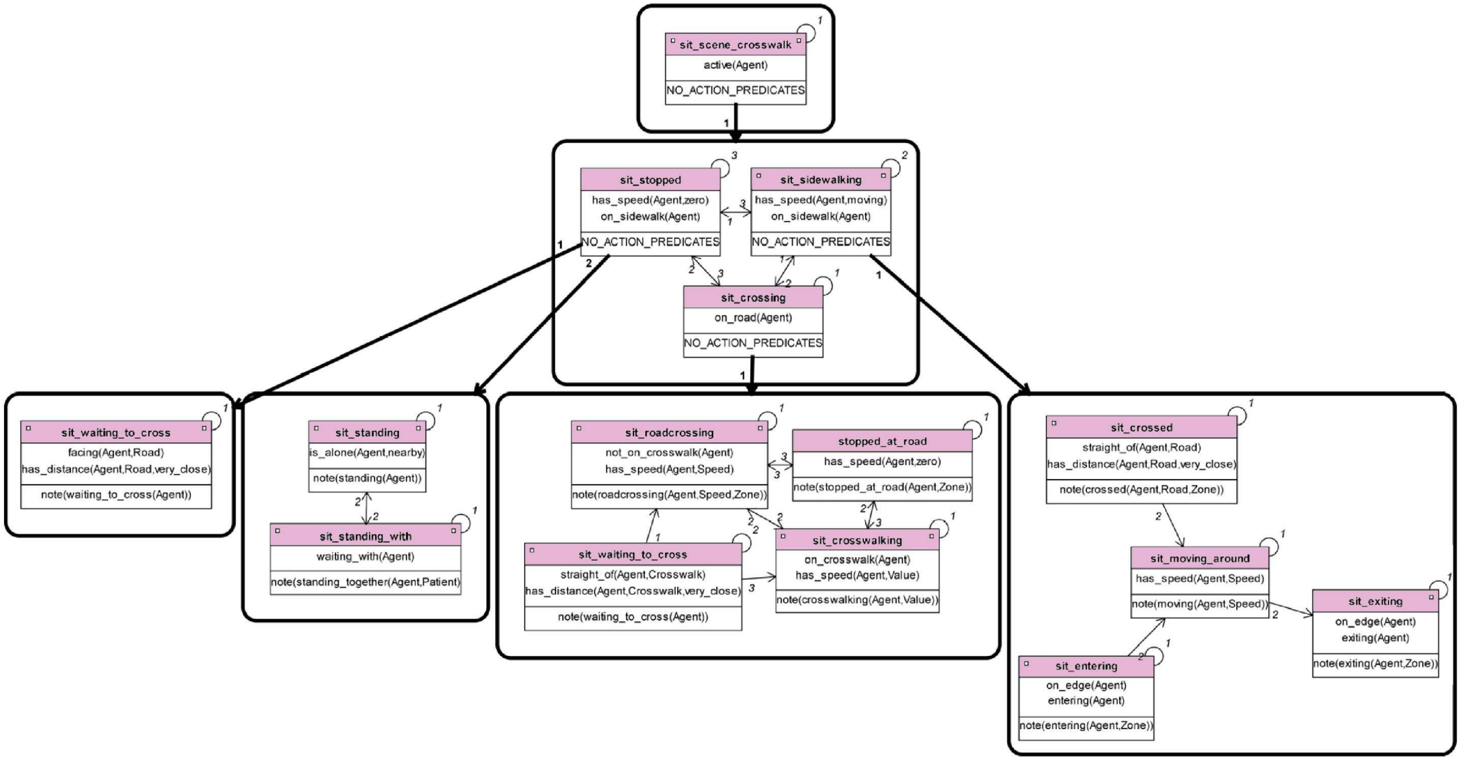


Fig. 3. Situation graph tree describing the behaviors of pedestrians on a crosswalk. Situation graphs are depicted as rounded rectangles, situation schemes are shown as normal rectangles. Bold arrows represent particularization edges, thin arrows stand for prediction edges, and $\frac{3}{4}$–circle arrows indicate self-predictions. Small rectangles to the left or to the right of the name of situation schemes mark that scheme as a start- or end-situation [1]. A SGT needs to focus on *behaviors* of the agents, while avoiding dependance to a particular scenario. The more this approach is achieved, the more flexible and scene-independent the SGT will be.

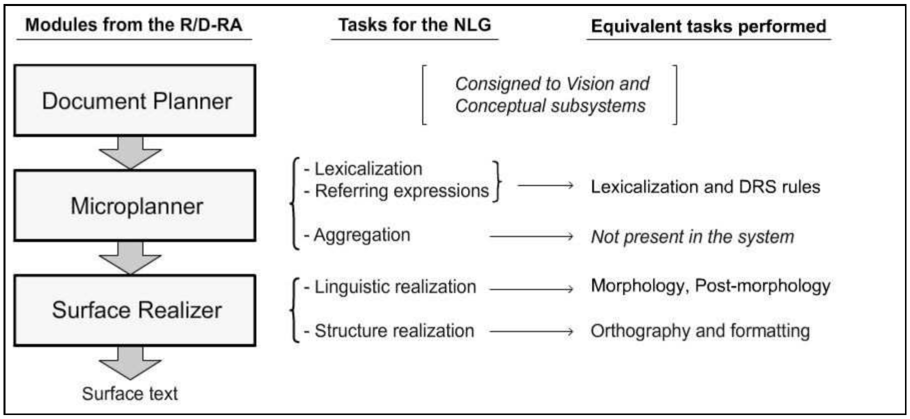


Fig. 4. Schema of Reiter/Dale Reference Architecture (R/D-RA) [8], including the tasks related to each module that are necessary for a Natural Language Generator.

Besides, the set of lemmata been used has to be extracted from a restricted corpus of the specific language. This corpus can be elaborated based upon the results of several psychophysical experiments on motion description, collected over a significative amount of native speakers of the target language. In our case, ten different people have independently contributed to the corpus with their own descriptions of the sample videos. Three different languages have been implemented for this scenario: Catalan, English, and Spanish.

3.1 Generation of textual descriptions

The overall process of generation of natural language descriptions is based on the architecture proposed by Reiter & Dale [8], which includes three modules; a document planner, a microplanner, and a surface realizer (see Fig. 4). The VS provides the information to be communicated to the user; this task is considered to be part of the Document Planner. The CS decides how this information needs to be structured and gives coherency to the results. This module provides general reasoning about the domain and determines the content to be included in the sentences to be generated, which are tasks related to the Document Planner, too. Further tasks, such as microplanning and surface realization, are included into the NS. An example for the entire process of generation is shown in Fig. 5.

Lexicalization. It is necessary to convert the abstract predicates from the CS into linguistic entities for communication, such as agents, patients, objects, or events, for instance. The classification of linguistically-perceived reality into thematic roles (e.g. agent, patient, location) is commonly used in contemporary linguistic-related applications as a possibility for the representation of semantics, and justifies the use of computational linguistics for describing content extracted by vision processes. The lexicalization step can be seen as a *mapping process*,

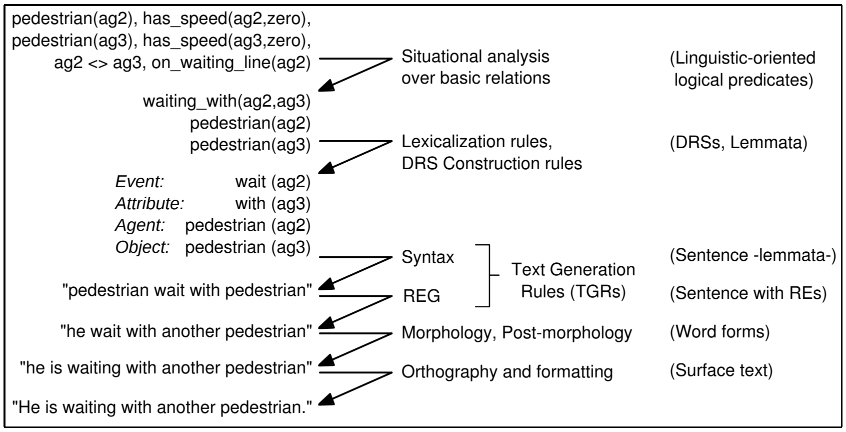

pedestrian(ag2), has_speed(ag2,zero),
pedestrian(ag3), has_speed(ag3,zero),
ag2 <> ag3, on_waiting_line(ag2)
Situational analysis over basic relations
(Linguistic-oriented logical predicates)
waiting_with(ag2,ag3)
pedestrian(ag2)
pedestrian(ag3)
Lexicalization rules, DRS Construction rules
(DRSs, Lemmata)
Event: wait (ag2)
Attribute: with (ag3)
Agent: pedestrian (ag2)
Object: pedestrian (ag3)
Syntax
REG
Text Generation Rules (TGRs)
(Sentence -lemmata-)
"pedestrian wait with pedestrian"
(Sentence with REs)
"he wait with another pedestrian"
Morphology, Post-morphology
(Word forms)
"he is waiting with another pedestrian"
Orthography and formatting
(Surface text)
"He is waiting with another pedestrian."

Fig. 5. Example for the generation of the sentence “He is waiting with another pedestrian” from logical predicates. The center column contains the tasks being performed, and the right column indicates the output obtained after each task.

in which the semantic concepts identifying different entities and events from the domain are attached to linguistic terms referring those formal realities. This way, this step works as a real dictionary, providing the required lemmata that will be a basis for describing the results using natural language.

Representation of the Discourse. Nevertheless, bridging the semantic gap between conceptual and linguistic knowledge cannot be achieved only with a lexicalization step. Discourse Representation Structures (DRSs) are the actual mechanism that facilitates to overcome the intrinsic vagueness of NL terms, by embedding semantics inferred at the conceptual level into the proper syntactical forms. Lemmata are just units that will be used by these structures to establish the interrelations which will convey the proper meaning to the sentences.

DRSs are semantic containers which relate referenced conceptual information to linguistic constructions [5]. A DRS always consist of a so-called *universe* of referents and a set of conditions, 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. Thus, the maintenance of the meaning for a discourse, including its cross-references, relations and cohesion can be granted, and linguistic mechanisms such as anaphoric pronominalization for referring expressions can be successfully implemented, e.g. *‘The pedestrian is running’* → *‘He is running’*. DRSs are included in both lexicalization and syntactical procedures, in form of Text Generation Rules (TGRs). A simple example for the tasks undertaken by a DRS is illustrated in Fig. 6.

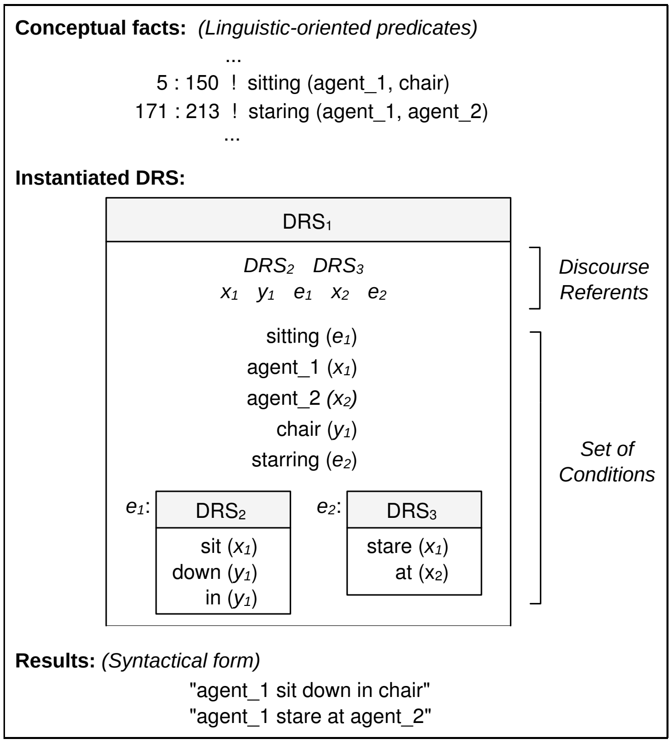

Conceptual facts: (Linguistic-oriented predicates)
...
5 : 150 ! sitting (agent_1, chair)
171 : 213 ! staring (agent_1, agent_2)
...
Instantiated DRS:
DRS1
DRS2 DRS3
x1 y1 e1 x2 e2
Discourse Referents
sitting (e1)
agent_1 (x1)
agent_2 (x2)
chair (y1)
starring (e2)
Set of Conditions
e1:
DRS2
sit (x1)
down (y1)
in (y1)
e2:
DRS3
stare (x1)
at (x2)
Results: (Syntactical form)
"agent_1 sit down in chair"
"agent_1 stare at agent_2"

Fig. 6. A pattern DRS allows to convert a stream of FMTHL into a string of textual symbols. In this particular case, the DRS is instantiated when `‘sitting(Agent, Location)’` or `‘staring(Agent, Patient)’` predicates are validated. Textual strings "agent_1 sit in chair" and "agent_1 stare at agent_2" would then be generated by the TGRs and forwarded within the NS.

Surface Realization. The Surface Realization stage is accomplished in two steps. A first morphological process applies over each single word and partially disambiguates the individual abstraction of that word, by means of morphological attributions such as gender or number. These attributions can be propagated upon the semantic relations previously established by DRSs among the lemmata of a single piece of discourse. After that, a set of post-morphological rules has been conceived to enable interactions among predefined configurations of words, thus affecting the final surface form of the text. This additional step is indispensable for many languages, in which certain phenomena force the surface form to change, e.g. contractions (*‘a’*+*‘el’*→*‘al’* in Spanish), or order variation (*‘es’*+*‘va’*+*‘en’*→*‘se’n va’* in Catalan).

4 Experimental Results

Some results for the situation analysis of the crosswalk scene are shown in Table 7. Textual descriptions in Catalan and Spanish have been selected for Agents 3 and 4, respectively. The English descriptions for these Agents are also included.

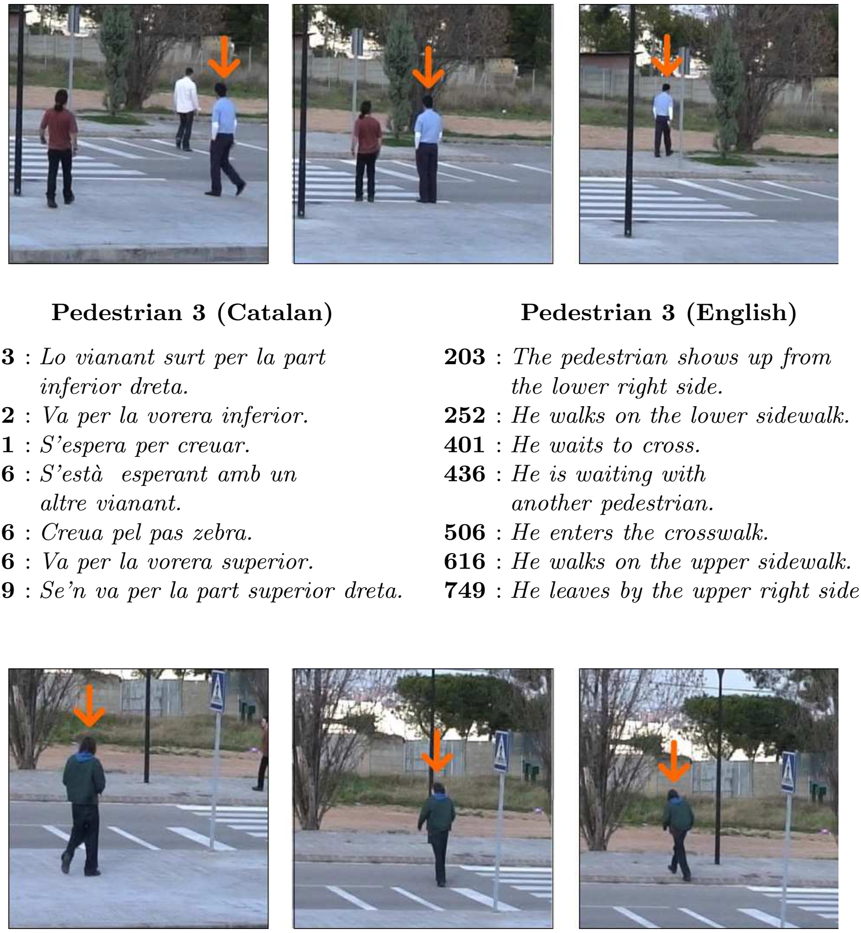
The natural language text generation was designed so that results matched perfectly with a selection of the descriptions provided by native users. The set of lemmata has been extracted from the different corpora contributed, the basic link between semantics and syntax has been provided by DRSs, and further grammatical considerations have been extracted from literature about the targeted languages. The Surface Realization stage has provided more flexibility for all type of language-specific morphological considerations, thus facilitating more natural results.

Language issues seem to be solved so far, at least considering the current capabilities of the CS. Next improvements should focus on the semantic content provided by the behavioral and inference subsystems, i.e. which situations must be considered for a certain domain and scenario, and in which way the reasonings for these situations have to be done.

5 Conclusions

A system that evaluates human sequences by generating natural language descriptions in multiple languages has been successfully developed in a first stage. It allows to automatically infer complex reasonings from the geometrical results extracted from vision processes in real-time. A deterministic approach has been chosen, applying methods based in Fuzzy Metric Temporal 'Horn' Logic, Situation Graph Trees, and Discourse Representation Theory. The conversion from quantitative information into qualitative conceptual predicates has been proved to be suitable for conceptual data manipulation and natural language generation.

The current NS allows for a flexible and fast incorporation of languages into a facility for multilingual generation of textual descriptions in natural language. The natural language formalism makes possible to generate fluid rich sentences to the user, allowing for detailed and refined expressions that are not possible by using other mechanisms. The interconnection of all the stages involved in the system has been proved as convenient for the whole process of evaluation, although several gaps still have to be solved. Further steps should include the extension of current behavioral models, the detection of groups and more complex interactions among agents and/or vehicles, and the use of uncertainty for not only predicting behaviors, but also inferring hypothesis of interpretation for the detected events within the scene.



Pedestrian 3 (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.*

Pedestrian 3 (English)

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

Pedestrian 4 (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.*

Pedestrian 4 (English)

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

Fig. 7. Some of the descriptions in natural language which have been generated for the crosswalk scene. The results match perfectly with the purposed set of natural language sentences suggested by a group of native speakers of the given languages.

Lastly, results from NL texts can be interpreted as semantic tags to provide content segmentation of the video sequences over time. We are currently studying the connection of a user interaction stage accepting input NL-based queries to a large database of video sequences, generic or specific. This will be the starting point for search engines capable of retrieving video sequences showing specific motion or factual contents. In addition to this, the segmentation of video sequences into time-intervals showing cohesive information can be applied for extracting a collection of few semantic shots from these sequences. This way, a compression of the relevant information – user-definable and freely configurable by declaring attentional factors – can be done by summarizing the entire videos with a list of behavior concepts. Thus, we aim to improve motion description patterns for video standards such as MPEG-7, thus allowing for high-level annotations related to the motion within the scene.

References

1. M. Arens and H.–H. Nagel. Representation of Behavioral Knowledge for Planning and Plan–Recognition in a Cognitive Vision System. Proc. of the 25th German Conference on AI (KI-2002), Aachen, Germany, pages 268–282. Springer 2002
2. H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. AI-magazine,78(1), 431–459, 1995. Elsevier Science.
3. J. Gonzàlez. Human Sequence Evaluation: The Key-Frame Approach. PhD thesis, Universitat Autònoma de Barcelona, Spain, 2004.
4. M. Haag, W. Theilmann, K. Schäfer and H.–H. Nagel. Integration of image sequence evaluation and FMTL programming. Proc. of the 21st Annual German Conference on AI: Advances in AI, 301–312. Springer-Verlag London, UK, 1997.
5. H. Kamp and U. Reyle. From Discourse to Logic. Kluwer Academic Publishers, Dordrecht; Boston, 1993.
6. A. Kojima and T. Tamura. 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.
7. H.–H. Nagel. Steps toward a Cognitive Vision System. AI-Magazine, 25(2):31-50, 2004.
8. E. Reiter and R. Dale. Building Natural Language Generation Systems. Cambridge University Press. Cambridge, UK, 2000.
9. D. Rowe, I. Rius, J. Gonzalez, and J.J. Villanueva. Improving Tracking by Handling Occlusions. Lecture Notes in Computer Science, in 3rd ICAPR, 2, 384–393. Springer. Bath, UK, 2005.
10. K. Schäfer and C. Brzoska. "F-Limette" fuzzy logic programming integrating metric temporal extensions. Journal of Symbolic Computation, 22(5-6):725–727. Academic Press, Inc. Duluth, MN, USA, 1996.