Spanish Text Generation for Image Sequence Evaluation using FMTHL and DRS

Carles Fernández*, Pau Baiget*, Mikhail Mozerov*, Jordi Gonzàlez+

* *Dept. dInformàtica, Computer Vision Centre, Edifici O. Campus UAB, 08193, Bellaterra, Spain*
E-mail: perno@cvc.uab.es
+ *Institut de Robòtica i Informàtica Ind. UPC, Llorens i Artigas 4-6, 08028, Barcelona, Spain*
E-mail: poal@cvc.uab.es

Abstract This contribution addresses the generation of natural language (NL) text descriptions for evaluation of human video sequences. The problem is tackled by converting initial geometrical information into Fuzzy Metric Temporal Logic predicates, which facilitates internal representations of the conceptual data and allow the temporal analysis of the situation by means of Situation Graph Trees. The results of the analysis are stored in Discourse Representation Structures, which will derive more easily into NL text generation in Spanish language.

Keywords: Natural language text generation, Discourse Representation Structure, Fuzzy Metric Temporal Horn Logic, Situation Graph Tree.

1 Introduction

The introduction of natural language interfaces into vision systems is becoming popular, specially in surveillance systems. Methods for describing human activities from video images have been described by Kojima and Tamura [7], and automatic visual surveillance systems for traffic applications have been studied by Nagel [8] or Buxton [3] among others. The considered system for evaluating human sequences is based on the Cognitive Vision System proposed by Nagel, which has been successfully used for automatic model-based evaluation of road-traffic videos.

In a visual surveillance system [10], 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 the most suitable predefined event that applies each case.

Natural language text generation for evaluation of human videosequences requires from three main disciplines, namely computer vision, knowledge representation and computational linguistics. Thus, the overall architecture of the system consists of three sub-systems [5] (see figure 1); a Vision sub-system (VS), which provides the geometric information extracted from a videosequence by means of detection and tracking processes, a Conceptual sub-system (CS), which infers the behavior of the agents from the conceptual primitives representing the scene facts, and a Natural Language sub-system (NS), which in principle comprises the natural language text generation for the output.

Discourse Representation Theory [4] seems to be of particular interest in this field, since it discusses algorithms for the translation of coherent natural language text into computer-internal representations by using logic predicates. The interpretation of tracked occurrences during the image evaluation provides the internal conceptual representation of the initial geometric results. These intermediate conceptual descriptions permit the analysis of the situations in the scene, which results are transformed into Discourse Representation Structures (DRS), and will be very useful in the derivation of natural language text.

The conversion of human actions and states over time into Fuzzy Metric Temporal Horn Logic (FMTHL) predicates [2] allows to reach the intermediate state, the conceptual representation layer, which facilitates the schematic representations of these scenarios. This inference system enables characterizations of uncertain and time-dependent data extracted from image sequences, and gives not only an interpretation for the behavior of the agent, but also reasonings for its possible reactions and predictions for its future actions [6].

The outline of the process includes two main differentiated procedures: in the first place, the Spanish language needs to be implemented in the NS subsystem. As stated in [5], no changes at the CS are necessary in order to extend the generation of descriptions in other languages. Besides this, the problem domain has to be redefined from traffic to human behaviors, which turns to be much more complicated due to the highly articulation of the human body and the independency of its parts. Only the main basic procedures are commented.

2 Implementation of Spanish

Angus2 is a program which is part of a system for the generation of natural language text (in English, German, Czech, and Japanese language) from video sequences [8]. Specifically, it implements a natural language subsystem, and part of a conceptual subsystem.

As has been described already in [5], there are at least four components that have to be written for each new language individually. These components are *lexicalization*, *text generation rules*, *morphological rules*, and minimal changes in *orthography*. Besides, the set of lemmas to be used has to be extracted from a restricted Spanish *corpus*. The next paragraphs briefly discuss how these components have been implemented for the Castilian language.

First of all, a small Spanish corpus has been built by 6 native speakers of the language. This corpus contains the minimum set of lemmas that are needed by the system to generate natural language sentences. Once the corpus has been built for the concrete implementation of a language, all the possible grammatical categories needed by the system can be established. The final chosen set of categories are shown in figure 2.

At this stage, the logical predicates imported from the Conceptual Layer are detected and clustered into appropriated lemmas. Two components are used during lexicalization: transformation rules and lexicalization rules.

Transformation rules do not vary from one language to the other (see [5]), while lexicalization rules

Grammatical category	Examples	Translation
Noun	*calzada, acera*	*lane, sidewalk*
Pronoun	*se*	*---*
Adjective	*superior, derecha*	*upper, right*
Determinative	*el, la*	*the*
Verb	*aparecer, cruzar*	*appear, cross*
Preposition	*por, para, junto a*	*by, for, next to*
Adverb	*peligrosamente*	*dangerously*

Figure 2: Accepted set of morphological categories that have been finally chosen to provide the lemmas for the text generation.

are language-dependent. The main aim of this second type of rules is the detection of a logic predicate and its conversion into a specific lemma in the target language, considering the order of the rules to apply. Nouns are complemented by adjectives at the beginning of the Lexicon, and verbs are extended with prepositions next, in order to get a simple syntax (although it can be artificial in some sense). Finally, some rules concerning adverbs were added for specific situations.

Text generation rules are also specific for each language. They infer the syntactical order of the input lemmas and inflect the morphological cases (number, gender, tense...). These rules are based on the detection and transformation of Discourse Representation Structures. The morphological and orthographical modifications have also been applied at this point.

3 Domain conversion

Spanish text generation currently provides good results in both of the given traffic situations. At this point, a step has to be done towards better approaching to the main line of the project, this is, the *natural text description of human sequences* in a given context. This new objective will be covered by generating correct Spanish text descriptions from an artificial situation, in which the inputs will be provided by hand a priori.

The chosen situation to work with has finally been a crosswalk scene (see figure 3). 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. One important

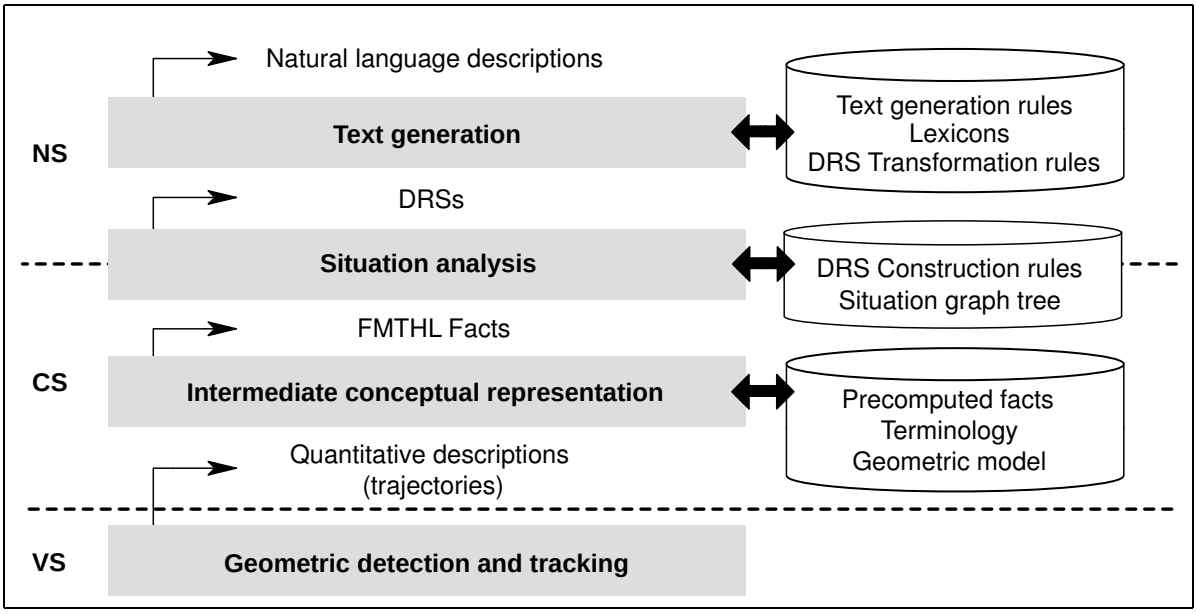

Figure 1: General schema of the stages and interfaces related to the generation of natural language text descriptions from human videosequences. The left acronyms represent different sub-systems that conform the whole system, the boxes describe the main processes that produce changes in data representations, and the right components specify some of the external tools required by the processes.

goal in this case will be the detection of dangerous behaviors, as for example crossing by the road and not by the crosswalk.

Figure 3: Original pedestrian crosswalk scene.

There are several needs to cover in order to create this new scene, which are described next.

3.1 Scene Modeling

The very first step has been to provide a well-designed scenario in which pedestrians can perform their actions. A geometrical modeling of the location has been done first in a groundplane bidimensional approach (figure 4), so a set of points, lines and sections are declared to distinguish the relevant topographic or interesting elements in the scene.

A second source of knowledge contains some more logical statements in order to confer conceptual meanings on the initial geometrical descriptors. This will be useful for identifying significative regions, so the movements and interactions of the agents can be contextualized.

3.2 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 over the agents[9].

Four agent trajectories have been built up. They consist on a set of Limette logical predicates of type `has_status`, in which a certain agent is related to a spatial position, orientation, speed and action tag for each specific instant of time. The kind of action is not considered at this stage of the project, since it is not necessary yet. Each considered trajectory represents a certain behavior of the agent.

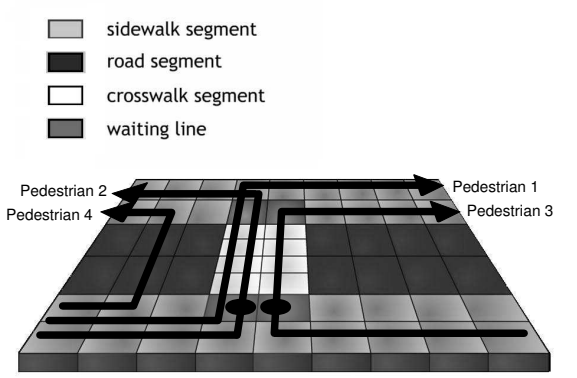

Figure 4: Groundplane schematic map of the main regions considered in the crosswalk scene. Pedestrian trajectories have been included. Black circles represent a stop in the waiting line.

3.3 Language models and terminology

The main difference that needs to be handled when switching from one problem domain to the other – traffic to human– lies in the language modeling for the terminology. Thus, most logical predicates need to be erased, modified or newly created for the situations dealing with the current language domain.

The terminology files contain the logical predicates needed to turn the quantitative values into qualitative knowledge for the specific domain. This knowledge is built up starting from the `has_status` information from the trajectories, and increases by developing new logical concepts concerning proximity, interaction, occupation or any other useful statement to evaluate the agent, offering qualitative results such as the perceptual distance between entities–*close*, *very close*, *far*–, the presence of obstacles in the path being followed by the agent, or the current kind of region in which the agent is positioned.

3.4 Corpus for the situation

A mandatory step refers to the writing of a corpus in every desired language, made by native speakers. It is necessary in order to obtain natural text generation in the output for the specific situation being evaluated.

The elaboration of the corpus can be made upon the results of several psychophysical experiments on motion description, taken over a significative amount of native speakers of the target language. In this case, six different people have independently contributed to the corpus with their own descriptions of the scene.

3.5 Situation graph tree (SGT)

The behavior of each agent will be represented in form of situation graph trees, which consist of situation schemes that describe the state of an agent and its environment at one discrete point of time, and the action that is supposedly carried out by the agent in that state. 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 on the SGT. It is necessary for it to have enough 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 *fuzzy metric temporal Horn logic* (FMTHL) for automatic exploitation of these behavior schemes. This means that the results of the evaluation sometimes turn to be nondeterministic and, in addition to this, there is no *a priori* method to assure the simplest and more efficient tree structure that better implements a solution to the problem.

A situation graph tree has been designed for the crosswalk scene (figure 5). This SGT has been traversed and translated into FMTHL knowledge

The obtained results for the situation analysis are shown in figure 6.

4 Conclusion

A system that evaluates human sequences by generating natural language descriptions in Spanish has been successfully developed in a first stage. A deterministic approach has been chosen, following the methods based in Fuzzy Metric Temporal Logic and Discourse Representation Structures used for traffic surveillance mechanisms. The conversion of quantitative information into qualitative conceptual predicates has been proved to be suitable for conceptual data manipulation and natural language generation.

Current results are acceptable for a first introduction in the problem, but there exists the impression

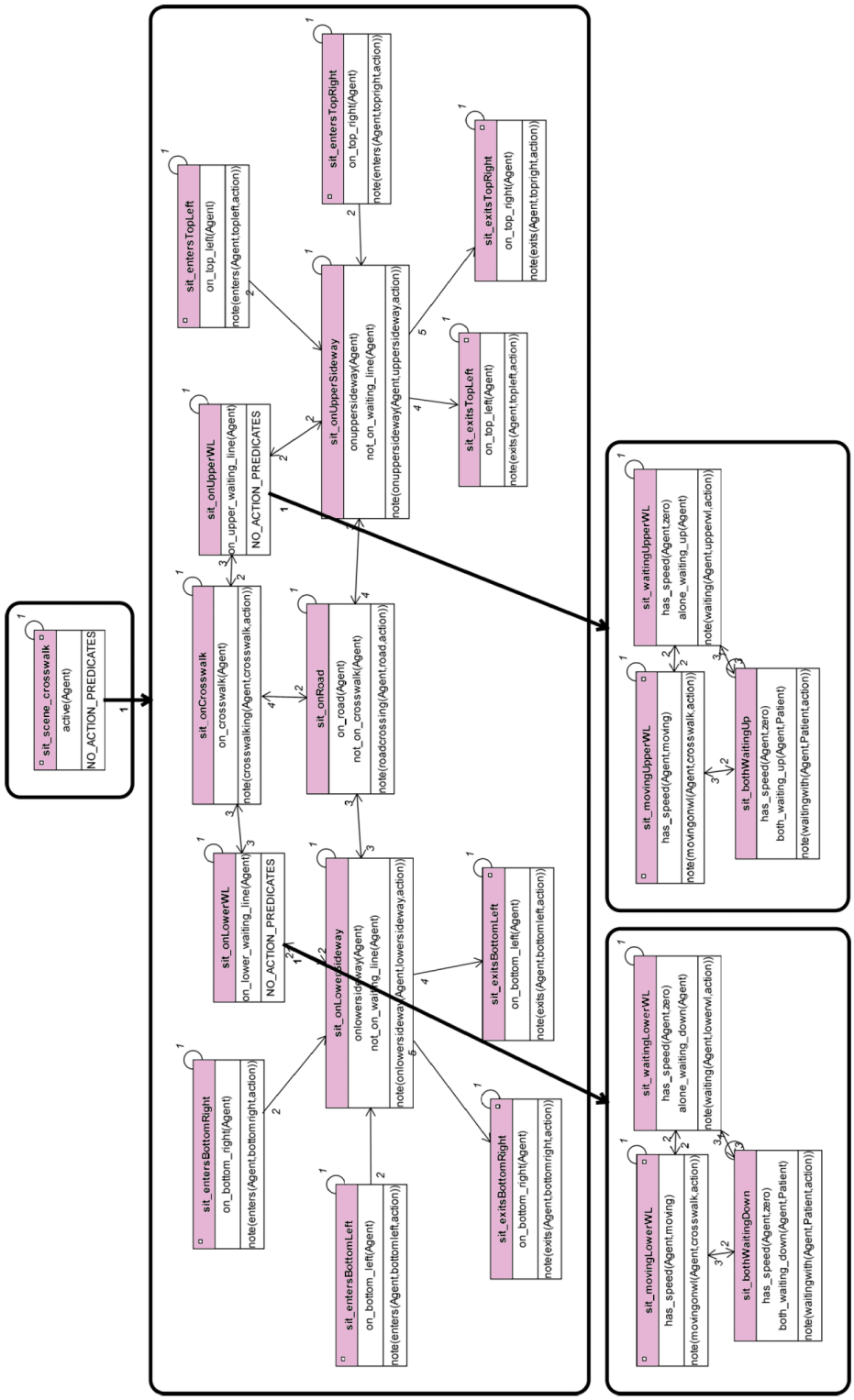

Figure 5: Situation graph tree describing the behaviors of pedestrians on a crosswalk. Situation graphs are depicted as rounded rectangles, and situation schemes are shown as normal rectangles. Bold arrows represent specialization edges, thin arrows stand for prediction edges and circle arrows indicate 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]

Pedestrian 1

3 : *El peatón aparece por la parte inferior izquierda.*
27 : *Camina por la acera inferior.*
226 : *Cruza por el paso de cebra.*
372 : *Camina por la acera superior.*
478 : *Se va por la parte superior derecha.*

Pedestrian 2

203 : *El peatón aparece por la parte inferior izquierda.*
230 : *Camina por la acera inferior.*
436 : *Está esperando junto a otro peatón.*
507 : *Cruza por el paso de cebra.*
652 : *Camina por la acera superior.*

Pedestrian 3

203 : *El peatón aparece por la parte inferior derecha.*
252 : *Camina por la acera inferior.*
401 : *Espera para cruzar.*
436 : *Está esperando junto a otro peatón.*
506 : *Cruza por el paso de cebra.*
616 : *Camina por la acera superior.*
749 : *Se va por la parte superior derecha.*

Pedestrian 4

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

Figure 6: Spanish descriptions generated for the four pedestrians considered in the crosswalk scene.

that the generated sentences do not reach a semantic level, but only a syntactical one. Further investigations have to exploit the resources offered by these conceptual and logic-oriented methods in order to increase the naturalness of the output text. The addition of other techniques regarding automatic learning implementation needs to be considered, too.

Acknowledgements

This work has been supported by EC grant IST-027110 for the HERMES project and by the Spanish MEC under projects TIC2003-08865 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, H.-H. Nagel. Representation of behavioral knowledge for planning and plan-recognition in a cognitive vision system. Proceedings of the 25th German Conference on AI, Springer, Aachen, Germany, 268–282, 2002.

[2] M. Arens, A. Ottlik and H.-H. Nagel. NL Texts for a Cognitive Vision System. ECAI2002, Proceedings of the 15th European Conference on AI, Lyon, July, 21–26, 2002.

[3] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. AI-magazine,78(1), 431–459, 1995. Elsevier Science.

[4] R. Gerber and H.-H. Nagel. (Mis?)-Using DRT for generation of NL text from image sequences. Lecture notes in computer science, 255–270, Springer.

[5] A. Fexa Dependence of conceptual representations for temporal developents in videosequences on a target language. Internal Report. IAKS der FI. Universität Karlsruhe. 2006.

[6] M. Haag, W. Theilmann, K. Schäfer and H.-H. Nagel. Integration of image sequence evaluation and FMTL programming. Proceedings of the 21st Annual German Conference on AI: Advances in AI, 301–312. Springer-Verlag London, UK, 1997.

[7] A. Kojima and T. Tamura. NL description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2), 171-184, 2002.

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

[9] D. Rowe, I. Rius, J. González, and J.J. Villanueva. Improving tracking by handling occlusions. In 3rd ICAPR, Bath, UK, volume 2, 384393. Springer LNCS 3687, 2005.

[10] M. Thonnat and N. Rota. Image understanding for visual surveillance applications. Proc. of 3rd Int. Workshop on Cooperative Distributed Vision, 51-82.