Automatic Learning of Conceptual Knowledge in Image Sequences for Human Behavior Interpretation

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

* *Computer Vision Center & Dept. de Ciències de la Computació, Edifici O, Campus UAB, 08193 Bellaterra, Spain*
+ *Institut de Robòtica i Informàtica Industrial (UPC – CSIC),Llorens i Artigas 4-6, 08028, Barcelona, Spain*

Abstract. This work describes an approach for the interpretation and explanation of human behavior in image sequences, within the context of a *Cognitive Vision System*. The information source is the geometrical data obtained by applying tracking algorithms to an image sequence, which is used to generate conceptual data. The spatial characteristics of the scene are automatically extracted from the resuling tracking trajectories obtained during a training period. Interpretation is achieved by means of a rule-based inference engine called *Fuzzy Metric Temporal Horn Logic* and a behavior modeling tool called *Situation Graph Tree*. These tools are used to generate conceptual descriptions which semantically describe observed behaviors.

1 Introduction

A classical problem in computer vision is the analysis of human behavior in observed scenes, where *behavior* refers to human agent trajectories which acquire a meaning in an specific scene. Results obtained in this research may benefit in the human-computer interaction and the video-surveillance domains.

Current motion understanding systems rely on numerical knowledge based on (i) the quantitative data obtained from tracking procedures and (ii) the geometrical properties of the scene[1],[7]. Therefore, this process is usually scene-dependent, and a-priori information of the spatial structure of the scene is required. The questions about the *what* and *why* can be answered by reasoning about the tracking data and transforming it to semantic predicates which relates each tracked agent with its environment. Common problems are the *semantic gap* which refers to the conceptual ambiguity between the image sequence and its possible interpretations, and *uncertainty*, which raises due to the impossibility of modeling all possible human behaviors.

In order to cope with the uncertainty aspects, integration can be learnt using a probabilistic framework: PCA and Mixtures of Gaussians [8], Belief Networks [6, 10] and Hidden Markov Models [3] provide examples. However, probabilistic approaches do not provide semantic explanation for observed agent behaviors.

Alternatively, Fuzzy Metric Temporal Horn Logic (FMTHL) also copes with the temporal and uncertainty aspects of integration in a goal-oriented manner [12]. This predicate logic language treats dynamic occurrences, uncertainties of the state estimation process, and intrinsic vagueness of conceptual terms in a unified manner. FMTHL uses three different strategies to accomplish such an abstraction process, according to the source of knowledge which is exploited to generate the qualitative description [12]. The main advantage of FMTHL over the previously referred algorithms relies on the promise to support not only the interpretation, but in addition diagnosis during the continuous development and test of the so-called *Cognitive Vision Systems* [9]. Further, Situation Graph Trees (SGTs) constitute a suitable behavior model which explicitly represents and combines the specialization, temporal, and semantic relationships of the constituent conceptual predicates in FMTHL. [4].

This contribution is structured as follows: next section describes the quantitative knowledge obtained from tracking. Next, we automatically build a conceptual scene model from the trajectories obtained from tracking during a training period. Consequently, contrary to other approaches, the geometrical properties of the scene are not provided beforehand, but automatically learnt from tracking instead. In Section 4 we show the conceptual description of human behaviors observed in a pedestrian crossing scene. Finally, Section 5 concludes the paper and shows future lines of research.

2 Numerical Knowledge for Motion Understanding

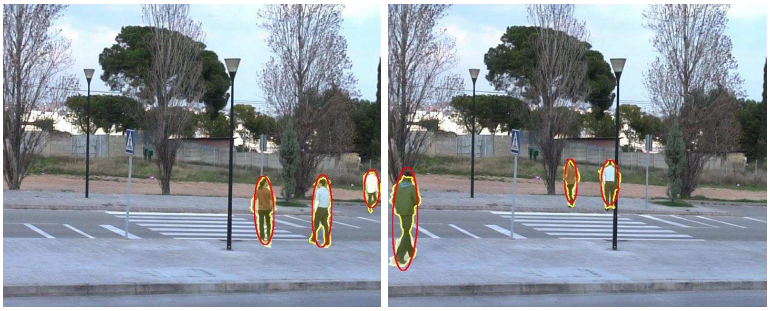
This section presents our procedure which converts the geometrical information obtained from tracking processes into a list of conceptual predicates which semantically describes motion events.

2.1 Information about the Agent

Semantic interpretation is based on the numerical *state vector of the agent*. The state vector is determined by the nature of the parameters used for tracking, which may refer to dynamical, positional and postural properties of the human agent. For example, motion verbs, such as accelerating could be instantiated by evaluating the history of the spatial and velocity parameters of the agent state. In our case, numerical knowledge obtained from tracking is comprised in the following attribute scheme:

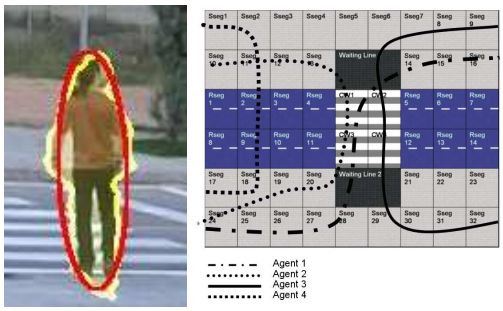
$$has\_status(Agent, pos, or, vel, aLabel),$$

which embeds the 2-D spatial position *pos* of the agent *Agent* in the floor plane, in addition to the velocity *vel* and the orientation *or*. These three parameters are called the *spatial status* of the agent. The parameter *aLabel* refers to the action, which obtained with respect to the velocity *vel*: we differentiate between *walking*, *standing* or *running*.



(a) **(b)**

Fig. 1. Example of human agent tracking in the pedestrian crossing scene.



(a) **(b)**

Fig. 2. Information of the scene. (a) The area occupied by an agent in a frame step is given by the envolving ellipse obtained in the segmentation process. (b) The roadway scene model and the agents' trajectories.

In our experiments, the quantitative description of the state of the agent is obtained through a segmentation process based on Horprasert algorithm [5] and on a subsequently state estimation method [11] for each time step, see Fig. 1.

2.2 Information about the Scene

Behavior analysis requires an explicit reference to a spatial context, i.e., a conceptual model of the scene. Such a model allows to infer the relationship of the agent with respect to (predefined) static objects of the scene, and to associate *facts* to specific locations within the scene. All this information is expressed as a set of logical predicates in FMTHL. The conceptual scene model is divided into polygonally bounded segments, which describe the possible positions in which an agent can be found.

Here we present a learning procedure for the conceptual scene model, which is based on the work of Fernyhough et al. [2]. We consider that the area (A)

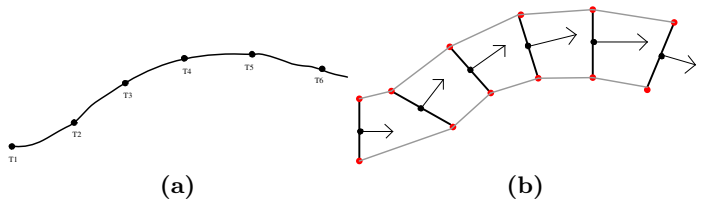


Fig. 3. Creation of trajectory segments. (a) Obtaining the temporal points of the trajectory. (b) Drawing lines perpendicular to the orientation of the agent for each temporal point, Connecting points at distance v from the corresponding temporal point, thus obtaining the trajectory segments.

occupied by an agent at each frame step is provided by the segmentation process, see Figure 2.(a). In our case, the trajectories of the agents were obtained from the pedestrian crossing scene shown in Figure 2.(b).

Initially, we obtain a set of temporal points of the trajectory (T_i), each one separated from the next by a fixed quantity of time, see Fig. 3.(a). Considering the orientation of the agent at each point T_i, a perpendicular line is drawn and the intersection (T_{i_r} and T_{i_l}) is found at distance A to the point T_i, see Figure 3.(b). Then, the T_{i_r} is joined with T_{i+1_r}, and T_{i_l} with T_{i+1_l}. Consequently, the four points T_{i_r}, T_{i+1_r}, T_{i_l}, T_{i+1_l} define a *trajectory segment*.

We build a matrix with the same height and width than the ground plane section of the scene. We traverse the matrix, assigning to each position (i, j) the number of trajectories for which one of their segments are drawn at the position (i, j) of the scene. For example, if three trajectories pass through the same point of the scene, this point's correspondence in the matrix will have value 3, see Fig. 4, where the brighter values represent the most accessed segments by the agents. Then, a threshold value is assigned depending on the number of trajectories analyzed and only those positions of the matrix whose value is equal to or exceeds this threshold are considered. Finally, all adjacent positions are considered to constitute a segment of the new scene, see Fig. 3.(c).

As a result of the learning process of the scene model, this is divided into polygonally bounded *roadway_segments*, which describe the positions in which an agent has been found. Each *roadway_segment*, has a label which determines the conceptual description associated with such a segment. At present, we manually distinguish (at least) four different types of segments, namely: *sideway_segment*, *waiting_line*, *roadway_segment*, and *crosswalk*. Consequently, we can build predicates which relate the spatial position of the agent with respect to these segments.

3 Working with Semantic Concepts

We next describe the use of FMTHL for generation of conceptual predicates from the state vector. Subsequently, we present a SGT for behavior modeling

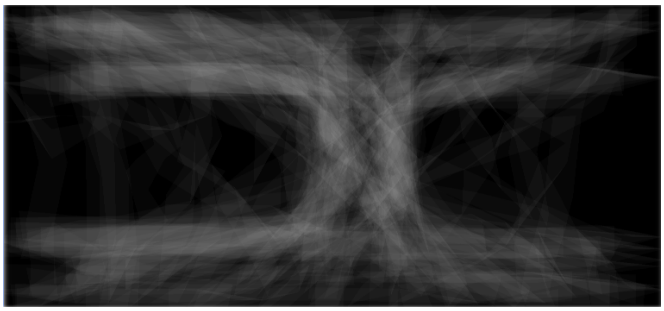


Fig. 4. Representation of the accumulation matrix used to obtain the segments of the new scene. Results match with the precomputed *roadway* scene.

which is used for the organization of plausible predicates into a temporal and conceptual hierarchy.

3.1 Generation of Conceptual Predicates

We use a system based on the Horn-Logic Fragment of *Fuzzy Metric Temporal Logic* (FMTHL)[12] which provides a suitable mechanism for processing fuzzy and temporal data. By means of a set of inference rules, FMTHL allows to transform quantitative data into qualitative data, expressed as logic predicates.

First, quantitative state parameters are associated to concepts like *moving*, *small*, *left*, or *briefly* with a fuzzy degree of validity characterizing how good a concept matches the numerical quantity. For example, the speed and orientation parameters of the state vector is associated to fuzzy attributes, thus allowing the instantiation of logic predicates such as $has\_speed(Agent, Value)$ or $has\_direction(Agent, Value)$.

Secondly, spatial relations are derived by considering the positions of the agents and other static objects in the scene. In this case, a conceptual scene model is required to describe the spatial coordinates of the agent with respect to static objects, other agents, and specific locations within the scene. This description is implemented by applying a distance function between the positions of different agents/objects in the scene. Subsequently, a discretization of the resulting distance value is obtained by using Fuzzy Logic, for example $is\_alone(Agent, Proximity)$ or $has\_distance(Agent, Patiens, Value)$.

Lastly, an action label is associated depending on the agent velocity. Thus, we can distinguish between three different actions, namely *running*, *walking* or *standing* by defining predicates such as $is\_performing(Agent, aLabel)$.

3.2 Situation Graph Trees

In this paper, predicate evaluation is performed in a goal-oriented manner: we use SGTs to recognize those situations which can be instantiated for an observed

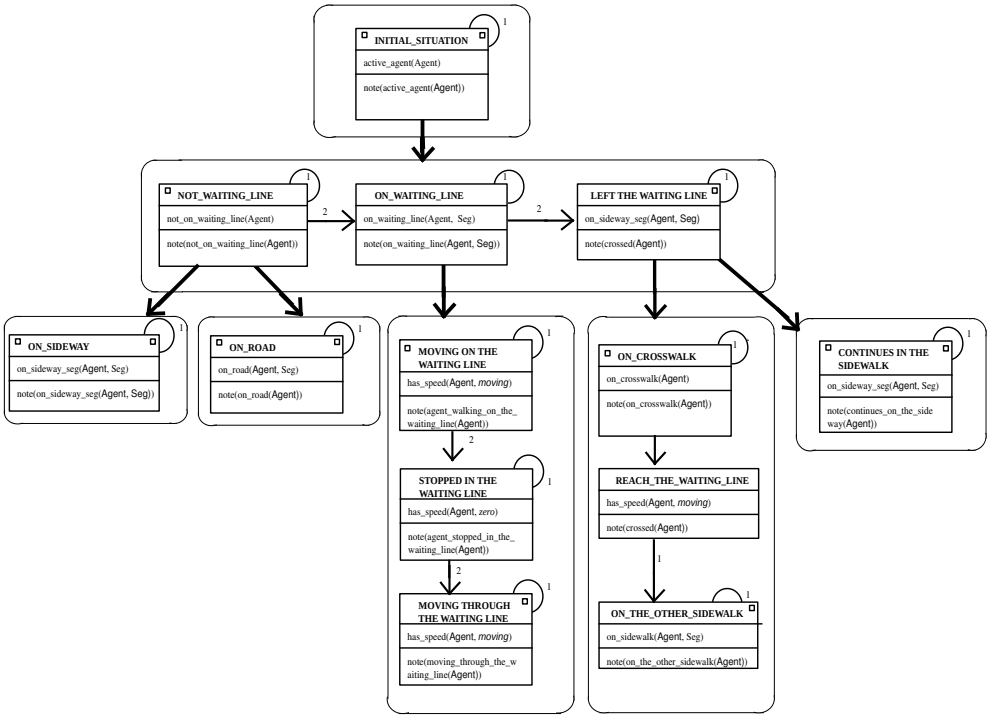


Fig. 5. Situation Graph Tree for behavior interpretation of human agents crossing a roadway.

agent by applying the so-called *graph traversal* [4]. The goal is to determine the most specialized situation which can be instantiated by considering the state vectors of the agents at each frame step. This traversal of the SGT is applied by considering the knowledge encoded in the form of prediction and specialization edges: on the one hand, given a situation, only its successors in temporal order will be evaluated at the next time step. On the other hand, each situation can be described in a conceptually more detailed way, thus allowing certain level of abstraction and specificity.

Fig. 5 depicts a simplified version of an SGT which allows to infer the behavior of agents of the roadway scene, as detailed next. The root graph comprises only one situation scheme, in which the predicate states that an agent is presently active, *active(Agent)*. The first possible specialization is the fact that the agent is not currently walking on the walking line. Then, only two situations can be instantiated: the agent is on the road or is on the sideway. Because in this scene there are only two kinds of segments where an agent can appear, this situation would repeat until the agent reaches the waiting line or it leaves the scene. When the agent arrives at the waiting line (*ON_WAITING_LINE*) the agent might stop for checking that there is no car on the road. This case is also modeled in the

Start	End	Situation
1	**26**	on_sideway_seg(**agent_1**, sseg24)
27	**76**	on_sideway_seg(**agent_1**, sseg25)
77	**126**	on_sideway_seg(**agent_1**, sseg26)
127	**179**	on_sideway_seg(**agent_1**, sseg27)
180	**184**	on_sideway_seg(**agent_1**, sseg28)
185	**225**	agent_walking on_waiting_line(**agent_1**)
226	**321**	on_crosswalk(**agent_1**)
322	**371**	crossed(**agent_1**)
372	**26**	on_the_other_sidewalk(**agent_1**)

(a)

Start	End	Situation
523	**571**	on_sideway_seg(**agent_4**, sseg17)
572	**595**	on_sideway_seg(**agent_4**, sseg18)
596	**635**	on_road(**agent_4**, rseg9)
636	**680**	on_road(**agent_4**, rseg2)
681	**740**	on_the_other_sidewalk(**agent_4**)

(b)

Fig. 6. Sequence of conceptual descriptions generated for : (a) *agent_1* (b)*agent_4*.

specialization of this situation scheme. After leaving the waiting line, the agent can walk on the pedestrian crossing (*ON_CROSSWALK*) or continue walking on the sideway. Once an agent has reached the sideway on the other side of the road, he or she is expected to continue walking on the sideway until leaving the scene.

4 Experimental Results

In this section we show the resulting process of predicate generation in order to obtain conceptual descriptions from image sequences recorded at a pedestrian crossing scenario.

The image sequence used for this purpose comprised 4 human behaviors, as summarized next:

- *Agent 1* walks on the sideway towards the waiting line and crosses the pedestrian crossing without stopping to see whether a car is approaching.
- *Agent 2* and *Agent 3* behave initially like *Agent 1*, but they stop on the waiting line for a few seconds before crossing the crosswalk.
- *Agent 4* crosses the road without going to the pedestrian crossing.

Figure 6 show the resulting conceptual descriptions generated for *agent_1* and *agent_4* using the SGT of the previous section. The resulting conceptual information establishes a behavioral layer in a cognitive vision system, which lets knowing what the agent is doing at each frame step and predict what the agent will probably do in the future frames. The first fact helps to generate natural language descriptions about the video sequence. The second allows the vision system to recover from segmentation errors e.g. predictable agent occlusions.

5 Conclusions

We have used a deterministic model suitable for modeling human behaviors. Our information source has been an image sequence previously processed with

pattern recognition algorithms, thus extracting quantitative data of the paths followed by human agents. Interpretation of human behavior has been achieved by means of a rule-based inference engine called FMTHL, and a human behavior modelling tool called Situation Graph Tree. This model has been tested for a street scene, and conceptual descriptions have been generated which semantically describe observed behavior.

At present, the SGT described here has not learning capabilities, so the accuracy of the modelled behavior will depend on the accuracy of the a-priory knowledge used. We also need to provide machine learning capabilites to improve reasoning through the sets of training examples. This will allow to confront sociological theories about observed human behavior, whose quantitative base is at present being computed from statistics and not from semantic concepts.

Acknowledgments. This work has been supported by the EC grant IST-027110 for the HERMES project and by the Spanish MEC under projects TIN2006-14606 and and DPI-2004-541. J. Gonzàlez acknowledges the support of a Juan de la Cierva postdoctoral fellowship from the Spanish MEC.

References

1. H. Buxton. Learning and understanding dynamic scene activity: A review. *Image and Vision Computing*, 21(1):125–136, 2002.
2. J. Fernyhough, A. Cohn, and D. Hogg. Constructing qualitative event models automatically from video input. *Image and Vision Computing*, 18:81– 103, 2000.
3. A. Galata, N. Johnson, and D. Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.
4. M. Haag and H.-H. Nagel. Incremental recognition of traffic situations from video image sequences. *Image and Vision Computing*, 18(2):137–153, 2000.
5. T. Horprasert, D. Harwood, and L. Davis. A Robust Background Subtraction and Shadow Detection. In *4th ACCV, Taipei, Taiwan*, volume 1, pages 983–988, 2000.
6. S.S. Intille and A.F. Bobick. Recognized planned, multiperson action. *International Journal of Computer Vision*, 81(3):414–445, 2001.
7. A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.
8. R.J. Morris and D.C. Hogg. Statistical models of object interaction. *International Journal of Computer Vision*, 37(2):209–215, 2000.
9. H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.
10. P. Remagnino, T. Tan, and K. Baker. Agent oriented annotation in model based visual surveillance. In *Proceedings of International Conference on Computer Vision (ICCV'98)*, pages 857–862, Mumbai, India, 1998.
11. D. Rowe, I. Rius, J. Gonzàlez, and J.J. Villanueva. Improving tracking by handling occlusions. In *3rd ICAPR, Bath, UK*, volume 2, pages 384–393. Springer LNCS 3687, 2005.
12. K. Schäfer. Fuzzy spatio-temporal logic programming. In C. Brzoska, editor, *Proceedings of 7th Workshop in Temporal and Non-Classical Logics – IJCAI'97*, pages 23–28, Nagoya, Japan, 1997.