Constructing a Path Database for Scene Categorization

Pau Baiget[*], Carles Fernández[*], Ariel Amato[*], Xavier Roca[*] and Jordi Gonzàlez[+]

[*] *Computer Vision Center, Dept. de Ciències de la Computació, Edifici O, Campus UAB, 08193 Bellaterra, Spain*
E-mail:pbaiget@cvc.uab.es

[+] *Institut de Robòtica i Informàtica Industrial (UPC – CSIC), Llorens i Artigas 4-6, 08028, Barcelona, Spain*

Abstract Human behavior understanding in image sequences requires to study human interaction with the environment, because human beings behave depending on their location. Therefore, a semantic labeling of the scenario must be performed and provided to the system. In this work we present a method to automatically extract semantic information from a scenario by means of a set of existing agent trajectories and an ontology of the possible semantic regions. Our training algorithm constructs a path database which serves to infer a conceptual model of the scenario and to predict future agent trajectories, even from different camera views.

Keywords: Human Sequence Evaluation, Cognitive Vision, Behavior Analysis.

1 Introduction

This work is focused on research towards the implementation of a cognitive vision system, capable of recognizing behavior patterns performed by human agents in an image sequence. In order to achieve this goal, we follow the *Human Sequence Evaluation* (HSE) scheme presented in [5]. Numerical data collected by motion trackers are discretized into a set of conceptual predicates, which allow to match the observed behavior with a set of predefined behavior patterns, and finally constructing a conceptual explanation of the image sequence [3]. Behavior analysis requires not only information about agents motion, but also information about their interaction with the environment. This is achieved by means of a semantically divided map of the scene, which must be provided to the system before the reasoning begins. Although an a–priori designed scene model is a good initial approach to obtain a suitable representation, see Fig. 1, an accurate observation of agent trajectories concludes that there could be a better distribution of regions that best fits with real human behavior, which is hard to define beforehand. Here we present a method that uses a set of agent trajectories to automatically divide the scene into segments and generate the conceptual scene model. The resulting scene representation is incorporated to a deterministic framework which uses (i) Fuzzy Metric Temporal Logic (FTML) to extract conceptual knowledge from numerical data collected from the image sequence [8, 9], and (ii) the Situation Graph Tree formalism to organize this knowledge into behavior patterns [1, 5]. This allows to extract conceptual explanations of complex behaviors and agent interaction, see [2] for details.

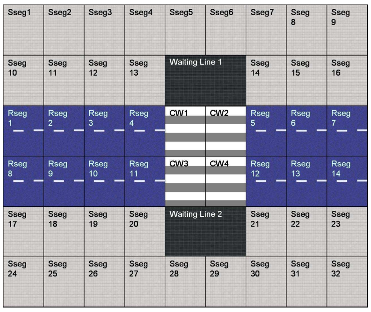
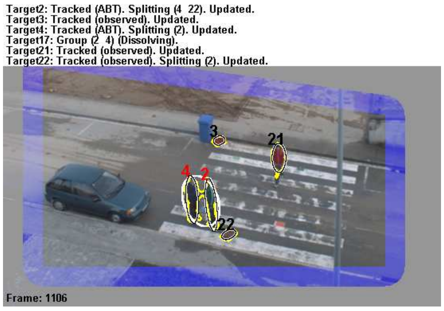


Figure 1: Predefined scene model for a pedestrian crossing scenario.

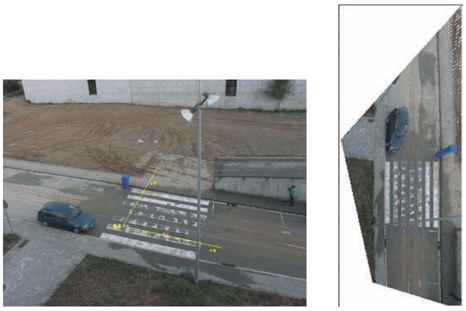
2 Identifying Semantic Regions

The creation of a conceptual scene model is divided into two steps: the learning step and the inference step. The learning step processes complete agent trajectories, represented in ground plane coordinates, and generates a database of common paths. The inference step classifies different regions of the scenario using a semantic label. Since semantic con-



```
Target2: Tracked (ABT). Splitting (4  22). Updated.
Target3: Tracked (observed). Updated.
Target4: Tracked (ABT). Splitting (2). Updated.
Target17: Group (2  4) (Dissolving).
Target21: Tracked (observed). Updated.
Target22: Tracked (observed). Splitting (2). Updated.
Frame: 1106
```

Figure 2: Trajectory acquisition from motion tracking



(a) (b)

Figure 3: Camera calibration. (a) Camera view (b) Ground plane coordinate view.

cepts can not be inferred only by training, this step requires to provide the system with an ontology describing the set of possible semantic labels in the scenariobe and rules defining their relations, see [4].

2.1 Tracking Architecture

Non-supervised multiple-target tracking involves such an inherent complexity that leads to propose a structured framework to accomplish such a task, see Fig. 2. This is implemented as a modular and hierarchically-organised system. The resulting architecture is based on a set of co-operating modules which are distributed through three levels. Each level is defined according to the different tasks to be performed: target detection, low-level tracking, and high-level tracking, see [6], [11],[10] for details.

At each time step, the tracker outputs agent position, orientation, speed, and its *spatial extent*, i.e the amount of area the agent occupies in camera coordinates. This information is further translated into ground plane coordinates [7] in order to obtain a representation independent from the camera view, see Fig. 3.

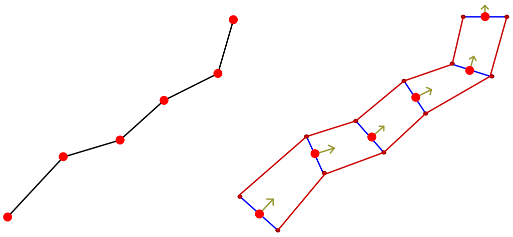


Figure 4: Obtaining paths from agent trajectories. More details in the text.

2.2 Path Database Construction

Our learning procedure constructs a database of common paths for a given scenario, represented in ground plane coordinates. It uses a wide set of existing agent trajectories, whose information includes the agent position, orientation and speed for each time step. Moreover, the *spatial extent* of the agent is provided, i.e the amount of area the agent occupies in the ground plane. The term *path* is defined in this paper as a portion of terrain which has been occupied by an agent during its trajectory. Figure 4 shows the conversion from trajectory to path. Considering the orientation O and the spatial extent A of the agent at each point T_i, a perpendicular line is drawn and T_{i_r} and T_{i_l} are found, both at distance A to the point T_i. Then, the point T_{i_r} is joined with the point T_{i+1_r}, the point T_{i_l} is joined with the point T_{i+1_l}, and the four points T_{i_r}, T_{i+1_r}, T_{i_l}, T_{i+1_l} define a portion of the resulting path. Each new trajectory is converted to a path and compared to existing paths in the database. Two paths are considered to be equivalent if they share a percentage of their respective area. In our experiments, we have found that a 85–90% of shared area is a good threshold to declare the equivalence of two paths. Figure 5 shows examples of equivalent and different paths. When two paths are found equivalent, they are merged into the database. If no equivalences have been found, the new path is added to the database. This database generation can be done online, creating new paths as new trajectories have been obtained from the vision system. Once the path database is considered stable, i.e. paths remain unchanged for a period of time, a threshold is applied in order to remove those paths having a low frequency in the training set, i.e. they are the result of

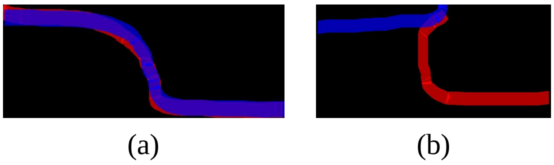

(a) (b)

Figure 5: Path comparison. Shared area between existing (blue) and the incoming path (red). (a) Equivalent paths (b) Non equivalent paths

very few merges during the training period. Remaining paths are considered to be the *common* paths the agents *usually* take in such a scenario.

2.3 Semantic Region Generation

The path database is analyzed in order to generate a conceptual model of the scene. The translation from quantitative data to qualitative concepts is done by means of an ontology [4], which declares the semantic labels that might be used to describe different parts of a scenario, and simple fuzzy logic predicates, which establish the typical relation between these labels. To describe this procedure, we use a pedestrian crossing scenario with *vehicle* and *human* agents, and available labels are *road*, *sideway*, *crosswalk and waiting line*.

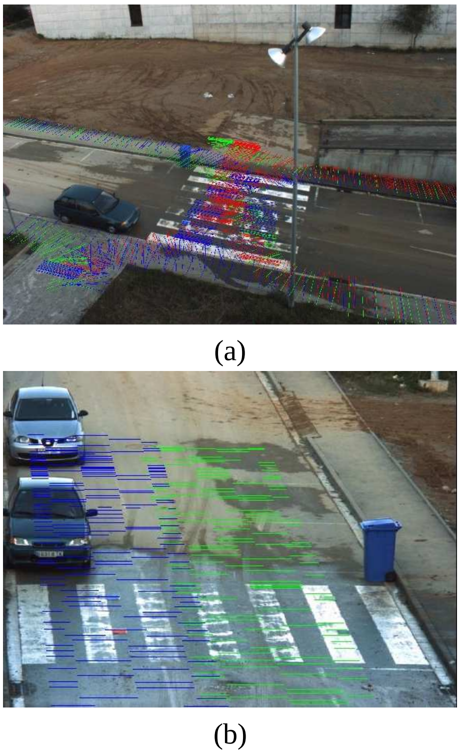
The conceptual scene model is constructed by means of *composite* and *leaf* regions. Each path of the database is considered to be a composite region. *Leaf regions* are determined by how do composite regions overlap and define a semantic portion of the scenario. Once composite and leaf regions have been obtained, they are labeled using a two–step inference process:

1. Each composite region is classified according to the kind of agent that performed the trajectories to construct the path.

```
road(CR):- vehicle_path(CR)
```

2. Leaf regions are labelled using logic rules that define the geographical relation between composite regions. For example, the following rule shows how to label a crosswalk region:

```
crosswalk(LR):- belongs(LR,CR),
         road(CR),
         belongs(LR, CR2),
         pedestrian(CR2)
```


(a)

(b)

Figure 6: Path database for (a) pedestrians (b) vehicles. Note that working with world coordinates eases the addition of new cameras, only demanding a previous calibration in order to use the obtained database.

3 Experimental results

In this section we show results obtained in the above mentioned procedures applied on image sequences recorded from an outdoor scenario, at a pedestrian crossing over a one–way road. The semantic labels used to classify regions are *road*, *sideway*, and *crosswalk*.

3.1 Scene Model Generation

In order to obtain a consistent region generation, we have used a wide trajectory set divided into two classes depending on which kind of agent, vehicle or pedestrian, has been tracked. This separation has been provided by the tracking process at the detection level and has been achieved by means of considering the size of the detected blob. Figure 6 shows the path databases obtained for pedestrians and vehicles, respectively.

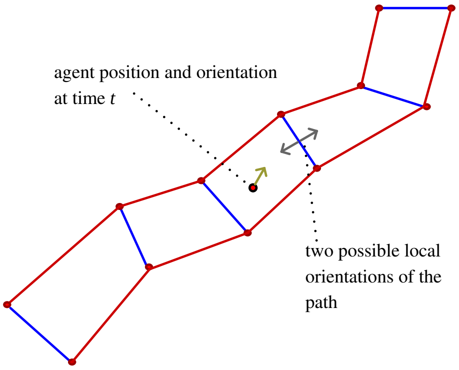


Figure 7: Path prediction. The current agent state (position and orientation) is confronted to each path of the database.

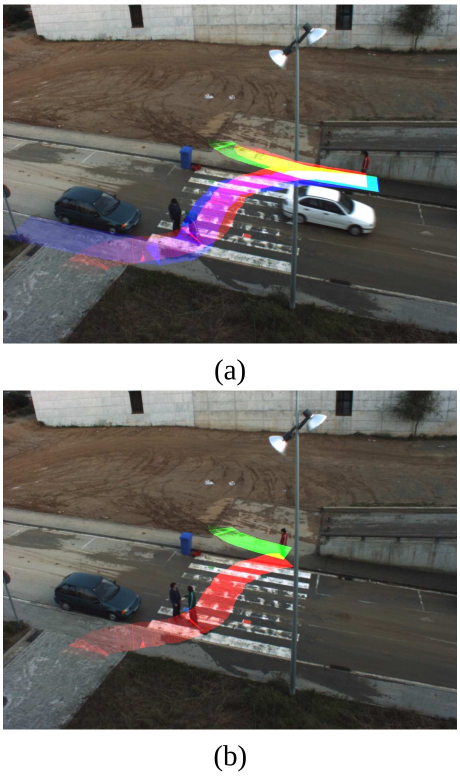
3.2 Applications

The results obtained by applying this method have two important applications in the HSE scheme. First, the path database obtained from the training period can be used to predict future agent trajectories, thus helping the tracking system. Second, the conceptual scene model inferred from the path database allows to reason about future agents interaction with their environment.

3.2.1 Agent Trajectory Prediction

When thresholded, the path database contains the *common* paths the agents follow in a given scenario. Hence, future agents being tracked in an image sequence are expected to follow one of the existing paths in the database. Although the database has been obtained by a set of existing trajectories and thereby it does not provide absolute truth value about future agent locations in the scenario, this information can be useful as a feedback to the tracking system introduced before.

Given a new agent in the scenario, the subset of paths containing the current agent position is selected, and those whose local orientation does match with the tracked agent orientation are shown as a probable paths for the agent in the further frame steps, see Fig. 7. Obviously, different predictions have a truth value according to the frequency of each path. The more frequent a path is, i.e. the number of trajectories merged to obtain this path, the higher the truth value is. Finally, the list of predictions is reported to the tracking architecture, which may use this information when it is unable to establish a position estimation, due to e.g. occlusions with the environment.


(a)

(b)

Figure 8: Results for agent trajectory prediction. More details in the text.

The training results have been applied to an image sequence and a complete trajectory prediction has been obtained. Figure 8 show possible paths that might be followed by the human agent wearing a red sweater. Each path has been colored choosing randomly one of the primitive colors (red, green or blue) and the color intensity at each pixel denotes the probability for the next agent position to be in that pixel.

3.2.2 Describing Agent Interaction with the Environment

The conceptual scene model obtained using the previous procedure can be further used to generate conceptual descriptions about incoming agent trajectories. We use the logic formalism Fuzzy Metric Temporal Logic (FMTL) to convert the quantitative data obtained from tracking, i.e. the agents positions for each time step, to qualitative concepts, which semantically describe the sequence of semantic regions the

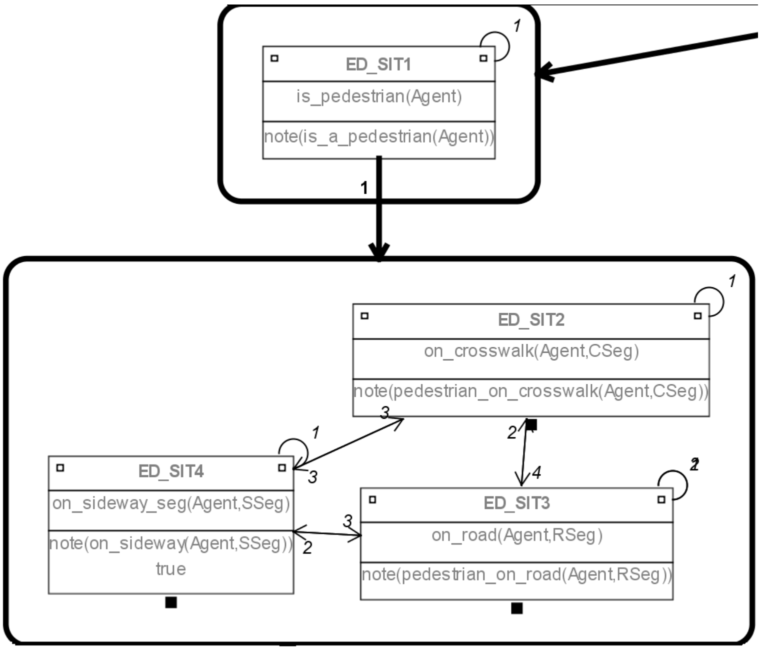

ED_SIT1
is_pedestrian(Agent)
note(is_a_pedestrian(Agent))
1
1
ED_SIT2
on_crosswalk(Agent,CSeg)
note(pedestrian_on_crosswalk(Agent,CSeg))
1
3
2
4
ED_SIT4
on_sideway_seg(Agent,SSeg)
note(on_sideway(Agent,SSeg))
true
1
3
ED_SIT3
on_road(Agent,RSeg)
note(pedestrian_on_road(Agent,RSeg))
2
3
2

Figure 9: SGT describing agent location in the pedestrian crossing scenario.

agent is walking through. Next, this list of predicates is used to match a predefined set of behavior patterns, being those defined using the Situation Graph Tree (SGT) formalism, see [1] for details. Fig. 9 shows the SGT used to describe the transitions between the possible semantic regions in the pedestrian crossing scenario. The predicates *on_road*, *on_sideway_seg*, and *on_crosswalk* are derived from the agent position at each time *t*, using the scene model. As an example, Table 1 shows a description obtained from a tracked human trajectory, which denotes an agent crossing the road while walking along the crosswalk.

4 Conclusions and Further Work

In this work we have presented a method to automatically extract conceptual information from a scenario by means of a set of existing agent trajectories, a calibrated camera, and an ontology of the possible semantic regions. Our training algorithm has constructed a path database which served to infer a semantic model of the scene and to predict future agent trajectories, even from a different calibrated camera. The resulting scene representation has been incorporated to a deterministic framework devoted to generate conceptual descriptions from a image sequence.

Start	End	Situation
1	105	on_sideway_seg_(**agent_1**,sseg_5).
106	225	on_sideway_seg(**agent_1**,sseg_6).
226	321	on_crosswalk(**agent_1**).
322	402	on_sideway_seg(**agent_1**,sseg_12).
403	524	on_sideway_seg(**agent_1**,sseg_13).

Table 1: Conceptual description of the regions crossed by an agent. This description denotes that the agent walked along the sideway during the frame interval 1–226 and then crossed the road using the crosswalk in the frame interval 226–321. Finally, the agent walked again in the sideway. Note that ensuring the agent crossed requires to distinguish between *left* and *right* sideway. This is easily achieved using FMTL rules like those explained in Section 2.3.

Future work will focus on improving the performance of the categorization of the scene. The first step is to take into account the temporal information, introduced as the third coordinate of the path shape. This will allow to differentiate paths depending on their temporal evolution. Moreover, trajectories will be processed online, and thus paths will be updated as the new agent positions are provided by the tracking system allowing the system to work with incomplete trajectories.

Acknowledgements

This work has been 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 Gonzlez 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. In *Proceedings of the 26th German Conference on Artificial Intelligence (KI-2003)*, pages 149–163. LNAI, Springler-Verlag: Berlin, Heidelberg, New York/NY, September 2003.

[2] P. Baiget, C. Fernández, X. Roca, and J. Gonzàlez. Automatic learning of conceptual knowledge for the interpretation of human behavior in video sequences. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007)*, Girona, Spain, 2007. Springer LNCS.

[3] P. Baiget, J. Gonzàlez, J. Orozco, and X. Roca. Interpretation of human motion in image sequences using situation graph trees. In *1st CVC Workshop on the Progress of Research and Development (CVCRD)*, volume ISBN 84-933652-8-9, Barcelona, Spain, 2006.

[4] C. Fernández, A. Fexa, and J. Gonzàlez. Ontologies for semantic integration in cognitive surveillance systems. In *2nd international conference on Semantics And digital Media Technology*, Genova, Italy, 2007.

[5] J. Gonzàlez. *Human Sequence Evaluation: The Key-frame Approach*. PhD thesis, Universitat Autònoma de Barcelona, Spain, 2004.

[6] I. Huerta, D. Rowe, M. Mozerov, and J. Gonzàlez. Improving background subtraction based on a casuistry of colour-motion segmentation problems. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (Ibpria 2007)*, volume 4477, pages 475–482, Girona, Spain, 2007. Springer LNCS.

[7] Mikhail Mozerov, Ariel Amato, Murad Al Haj, and Jordi Gonzàlez. A simple method of multiple camera calibration for the joint top view projection. In *5th International Conference on Computer Recognition Systems (CORES'2007)*, Wroclaw, Poland, 2007.

[8] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.

[9] H.-H. Nagel. Image sequence evaluation: 30 years and still going strong. In A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquézar, O. Eklundh, and Y. Aloimonos, editors, *Proceedings of International Conference on Pattern Recognition (ICPR'2000)*, volume 1, pages 149–158, Barcelona, Spain, 2000.

[10] D. Rowe, J. Gonzàlez, I. Huerta, and J.J. Villanueva. On reasoning over tracking events. In *15th SCIA, Aalborg, Denmark*, pages 502–511. Springer LNCS, 2007.

[11] D. Rowe, I. Reid, J. Gonzàlez, and J. Villanueva. Unconstrained multiple-people tracking. In *28th DAGM, Berlin, Germany*, volume LNCS 4174, pages 505–514. Springer, 2006.