Observing Human Behavior in Image Sequences: the *Video–Hermeneutics* Challenge

Pau Baiget, Jordi Gonzàlez

Computer Vision Center, Dept. de Ciències de la Computació, Edifici O, Campus UAB, 08193 Bellaterra, Spain
pbaiget@cvc.uab.es

euCognition Network Action NA - 149 – 1
www.eucognition.org

Abstract This work summarizes the current challenges in research towards learning and recognizing behavior in image sequences. The last decade has been focused on the acquisition of quantitative data in sequences, specially, the segmentation and tracking of moving targets. Although there still exist several open problems in this low level techniques, recent achievements in that areas have led to think of a next step: the semantically evaluation of image sequence contents, also called *video–hermeneutics*. In this paper we focus on the analysis of image sequences in video surveillance contexts and we consider human beings as the specific target, since this kind of target comprises all the challenges given the wide set of plausible behaviors to be analyzed.

Keywords: Motion Analysis and Recognition, Scene Learning, Behavior Understanding.

1 Introduction

In recent years, the evaluation of human behavior from image sequences, called *video–hermeneutics* in [10], has raised as a challenging approach to computer science: impressive developments have been possible thanks to a large number of technological advances in the hardware field. Emerging capabilities have led to a wide range of scientific contribution, and, subsequently, to new software implementations. Nowadays, huge amounts of data collected from vision systems are being analyzed to recognize and *understand* human behavior patterns. This goal

requires a high-level reasoning process to convert the quantitative data obtained from low-level vision systems into qualitative statements which semantically represent observed behavior. However, this task is complex due to several reasons. On the one hand, human motion is highly non-linear, a-priori unknown, and it is always subject to sudden and unforeseeable changes. On the other hand, human behavior depends on numerous factors like psichological ones —such as mood and culture— or physics ones —such as age and gender. Moreover, knowledge acquisition from image sequences involves two steps which can entail a degree of uncertainty to the image sequence interpretation. On the one hand, the error pulled from the computer vision subsystem caused by the *sensory gap*, which refers to the lack of accuracy in the quantitative data acquisition from the image sequences. On the other hand, human behavior modeling has to deal with the uncertainty due to the *semantic gap*, which refers to the conceptual ambiguity between the image sequence and its possible interpretations.

The remaining of the paper is structured as follows: next section overviews the problem of recognizing and explaining behavior in image sequences and reviews the two main existing approaches, namely Top–Down and Bottom–Up. Next, the automatic modeling of semantic scene models is explained. In Section 4 we introduce ontologies as a way to join Top–Down and Bottom–Up approaches. Finally, last section concludes the paper and shows future lines of research.

2 Behavior Recognition in Image Sequences

The analysis of behavior in image sequences consists of three steps: (i) obtain quantitative data from a computer vision system (ii) extract qualitative information, and (iii) interpret and classify into one behavior pattern. These behavior patterns can be either predefined (Top–Down) or learnt from previous observations (Bottom–Up). Next, we describe the two approaches and we summarize the advantages and drawbacks of each solution.

2.1 Quantitative Data from Computer Vision

This section reviews different computer vision results which are normally used to extract semantic knowledge about human behavior. Those results consist of a set of features captured for an agent frame by frame during its existence within the image sequence. Such systems are called *trackers* and maintain a temporal continuity and coherence between frames. Depending on the region of interest, tracking outputs can be joint into three main groups:

- *Agent Level*, describing the development of agent location and size, in image plane coordinates. The feature vector obtained for each frame usually

Figure 1: Computer Vision output. (a) Trajectory level (b) Body level (c) Face level.

includes the agent position (x, y) and a bounding box or ellipse estimating the spatial occupancy of the agent.

- *Body Level*, describing the movements of body parts over time.
- *Face Level*, describing the facial actions performed over time, usually represented by movements of different face attributes.

2.2 Behavior as a combination of interactions

In the context of Cognitive Vision, a *behavior* is considered as a sequence of actions performed by moving targets, which will be also addressed in this paper as *agents*, in an image sequence. These actions represent a change in the current state of the agent and are based on three types of interactions:

- *Interaction with the scenario*: agent behavior is explained depending on where the agent is located. The same action sequence can represent different behaviors in different locations. Hence, some prior knowledge about the scenario must be provided in order to disambiguate that situations. Different approaches will be sketched in further sections.
- *Interaction with static objects*: agents can interact with objects that initially formed part of the background, e.g. bags, vehicles, chairs, etc. Moreover, those objects could have been left by other agents some frames before. The interaction with each of those objects may represent a completely different behavior.
- *Interaction with other agents*: human behavior is determined not only by its own goals, but also by the reactions to other agents' behavior. Given that, reasoning about one agent implies reasoning about all agents at the same time. We can find lots of a–priori unknown relationships between agents, such as friendship, that might interlace the explanation of separate behaviors.

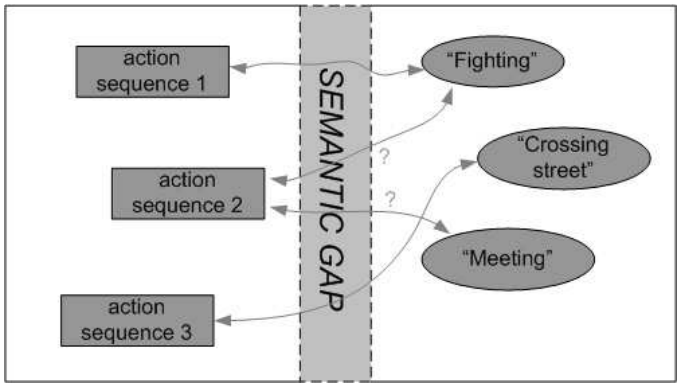


Figure 2: The semantic gap between the observations (quantitative) and the concepts (qualitative).

2.3 The semantic gap

Every system devoted to extract qualitative information from numerical data has to deal with a semantic gap. Roughly speaking, a computer does not understand semantic terms, i.e. those terms that have been established by human beings with respect to their representation in natural language.

In our context, the semantic gap is present in both learning and recognition of behaviors, see Fig. 2. On the one hand, although a computer can learn a set of behavior patterns given some training data, it is unable to extract a semantic explanation of the results. Indeed, it is not possible to assert if some of the learnt patterns represent a *normal* or an *anomalous* behavior unless the training set has been previously labeled. On the other hand, behavior recognition consists of matching an observed action sequence with a set of behavior patterns. In this case, the inaccuracy pulled from the vision system and the ambiguity of possible interpretations, see Fig. 3, can lead to a misclassification of the input action sequence.

2.4 Predefined Behavior Models

In order to cope with these drawbacks, some recent contributions have proposed the use of ontologies as a way to restrict the domain [7]. Furthermore, previous works have presented different approaches to represent the domain in order to increment the accuracy in behavior recognition. Thus, several models have been presented to define the domain of possible behaviors to be recognized in a selected environment. To cite few, Arens and Nagel [1] proposes a framework to combine the Situation Graph Tree structure (SGT) with a fuzzy temporal logic formalism to generate descriptions of observed human behaviors. Ghanem et al. [9] applies

Figure 3: Example of semantic gap. Given the gray agent trajectory, two possible interpretations can be done 1) The agent is searching his parked car (considered as *normal* behavior), and 2) The agent is trying to steal a car (*suspicious* behavior). Given that situation, no possible reactions can be done until further information is obtained or provided beforehand.

Petri Nets to model expected human behavior and Bremond et al. [5] uses a symbolic network combined with rule–based temporal constraints. An example of the previously mentioned structures is shown in Fig. 4.

Behavior understanding systems that make use of predefined behavior models show good performance recognizing the modeled behaviors. Moreover, these methods can provide a semantic explanation of the recognized behaviors, allowing to generate a natural language description of the image sequence [6]. Furthermore, since most of these approaches use fuzzy predicates to represent knowledge, uncertainty can be managed and incorporated to the system.

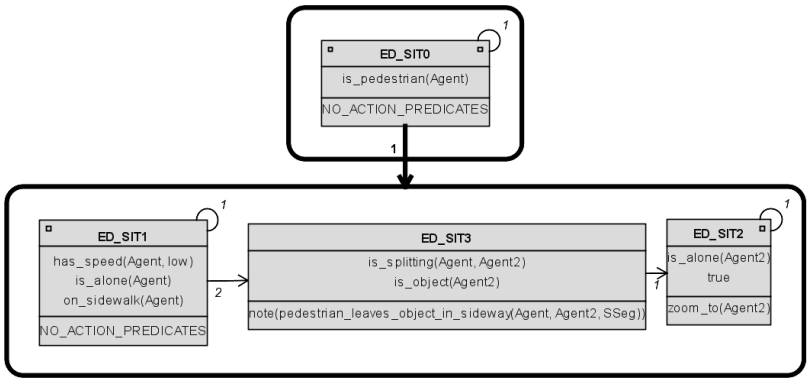
Nevertheless, top–down approaches depend on the modeling accuracy of human experts. Also, these methods do neither evolve over time nor learn from new observations, thereby being affected by any modification in the scenario or the incorporation of new behaviors.

2.5 Trajectory Based Behavior Learning

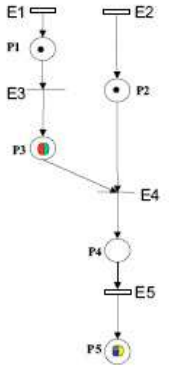
The process of learning behavior from the output of a vision system is the key topic in the research in the area of cognitive vision systems. Since designing predefined behavior models is a tough task and it has to be repeated for each new discourse domain, a system capable of learning and evolving over time has become a very attractive goal nowadays.

The common approach towards learning behaviors consists of training a model (HMM, DBN, SVM) by analyzing and clustering the outputs described in Section 2.1. However, the only level that is providing enough accuracy to enable such learning is the *Agent level*. Therefore, most of the recent approaches consider learning using the output obtained by motion tracking algorithms (trajectories):

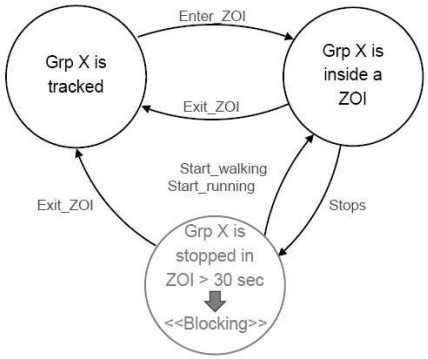
1. Agent position (x, y) over the image plane.



(a)



(b)



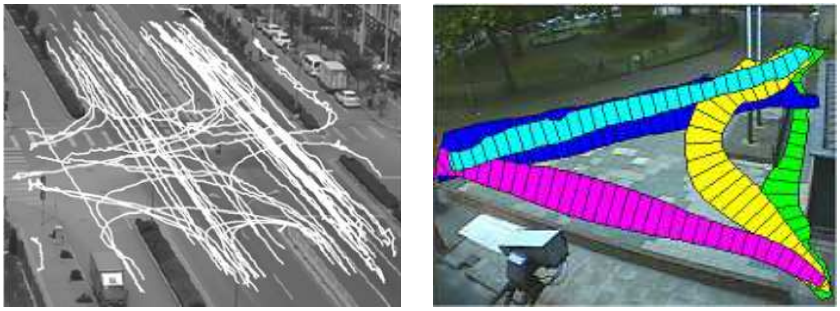
(c)

Figure 4: Predefined behavior models. (a) Situation Graph Tree [1] (b) Petri Nets [9] (c) Symbolic Network [5].

2. Bounding box or bounding ellipse of the agent.

3. Speed and orientation over the image plane.

The training set consists of a collection of trajectories that are clustered in order to generate *route models* which can be used to perform abnormal behavior detection or to estimate the agent position in further frame steps. The first approach was introduced by Johnson and Hogg [12], which statistically models the spatial distribution of trajectories using vector quantization. New trajectories are represented as sequences of vectors and are clustered using two competitive learning neural networks, one that finds the sequence of vectors that best represent a trajectory and the second to cluster those sequences. Stauffer and Grimson [17] use again vector quantization, but the clusters are identified by a hierarchical analysis of the vector co-occurrences in the trajectories. More recent approaches consider different representations such as offline [13, 11] and on–line [15] path modeling



(a) (b)

Figure 5: Learning route models. (a) Results from [11]. (b) Results from [13].

based on spatial extension of trajectories, B–splines [4], and HMM [16]. Fig. 5 shows obtained route models using different approaches.

The resulting route models represent the typical paths that agents usually take within the scenario. Obviously, the accuracy of the generated models depends on the representativity of the training data. Possible applications include:

- Anomalous behavior detection: Since training data are quantitative and no semantic knowledge is provided, new trajectories can only be classified into normal or abnormal behavior, depending on its similarity to the generated route models. Since learnt behavior patterns are based not on semantic concepts but on quantitative data, no semantic explanation of behavior can be provided unless the obtained route models are finally labeled [14].

- Trajectory prediction: Given a new agent inside the scenario, its current trajectory is continuously compared with the route models in order to estimate future agent locations and the most probable exit point from the scenario.

3 Learning Scene Features from Observations

As stated in previous sections, the representation of the scenario is crucial when reasoning about an observed action sequence. The most widely used representation is to divide the scenario in *semantic regions*, i.e. small regions with a semantic label. For instance, in an urban outdoor scenario we can find *crosswalk*, *sidewalk*, *road*, etc. However, there is a significant difference in the scene model design when considering either outdoor or indoor scenarios. While in the former, regions normally represent parts of the assumed ground–plane, in the latter there may be static objects the agents may interact to, e.g. chairs or doors.

The automatic acquisition of semantic knowledge about the scenario has recently received a wide interest from the research community. Fernyhough et al.

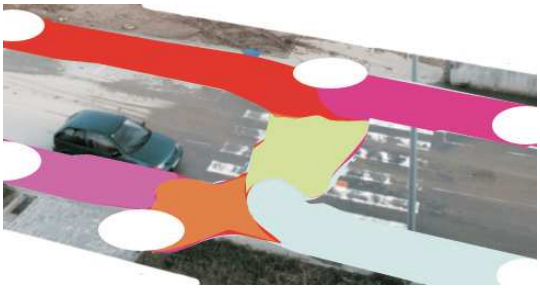


Figure 6: Learnt scene model after applying the approach in [3] to a pedestrian crossing scenario.

[8] proposed a method to learn and classify semantic regions from a scenario. This approach recognizes common paths by extending trajectories with the spatial extent occupied by the agents in camera coordinates. Although the method does not need any a–priori information, it requires full trajectories and cannot handle on–line learning. In addition, this method does not use orientation to compute paths and thus does not distinguish between objects following the same route but different directions. As an extension of this approach, Baiget et al. [2] applies a similar method to human domain using ground plane coordinates instead of camera coordinates. Despite showing promising intermediate results, the formulation lacks extensibility to other domains and no semantic description is reported. This lack of conceptual labelling of the scenario is addressed by Makris and Ellis in [14], learning entry/exit zones and routes from trajectory samples. The start/end points of trajectories are used to learn entry/exit zones applying the EM algorithm. Nevertheless, the method requires complete trajectories and the learning process is done offline. More recently, Baiget et al. [3] have proposed an on–line trajectory clustering method that organizes clusters to obtain a topological map of the scenario, which can be semantically labeled by using a domain ontology, see Fig 6.

4 Joining Top–Down and Bottom–Up Approaches

Nowadays, it seems that the most suitable mechanism that allows to maintain a semantic layer on a learning system are ontologies [7]. An ontology describes a set of concepts and their relations, telling the system the domain to be learnt and how to relate quantitative data with semantic concepts. However, in every domain the quantity of concepts and behaviors related to human agents is huge. Therefore, the ontology should be designed over a very restricted domain, e.g. *'surveillance on urban intersections'* or *'elderly care'*.

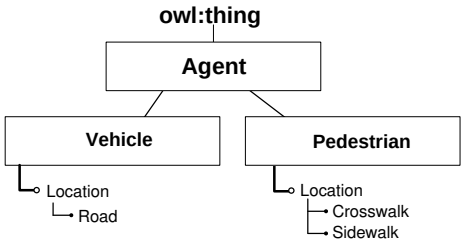


Figure 7: Ontology used to assign semantic labels to learnt zones.

	Top–Down	Bottom–Up
Learning	No	Yes
Reaction to new behaviors	No	Yes
Recognition of complex Behaviors	Depends on user	Depends on training
Semantic Description	Yes	No
Natural Language explanation	Yes	No

Table 1: Comparison of main features between Top–Down and Bottom–Up approaches.

For instance, consider a pedestrian crossing environment. The scene model depicted in Fig. 6 has been obtained by clustering pedestrian trajectories. All the resulting regions are labeled as *pedestrian_region*. Moreover, the same process is applied by processing vehicle trajectories and again the obtained regions are labeled as *vehicle_region*. The ontology in Fig. 7 describes which semantic labels can be found in such a environment. Given that, the matching between regions and semantic labels is inferred as follows:

$$\begin{aligned} road &\leftarrow vehicle\_region \\ crosswalk &\leftarrow road \cap pedestrian\_region \\ sidewalk &\leftarrow pedestrian\_region - crosswalk \end{aligned}$$

5 Conclusions

This work has presented the current open problems towards learning and recognizing behavior in image sequences. We have focused on the analysis of image sequences in surveillance contexts considering human beings as the specific target, since this kind of target comprises all the challenges given the wide set of plausible behaviors to be analyzed.

The most challenging question for future work is how to join Top–Down and Bottom–Up approaches in a way that all the advantages are kept and the draw-

backs are minimized, see Table 5. As stated in previous section, the use of ontologies seems to be an initial bridge to cope with the semantic gap between the qualitative concepts and the quantitative observations.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and euCognition:The European Network for the Advancement of Artificial Cognitive Systems (FP6-26408).

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, C. Fernández, X. Roca, and J. González. Dynamic scene conceptualization by trajectory clustering. *Computer Vision and Image Understanding, Special Issue on Intelligent Vision Systems*, 2008.

[4] P. Baiget, E. Sommerlade, I. Reid, and J. Gonzàlez. Finding prototypes to estimate trajectory development in outdoor scenarios. In *Proceedings of the 1st THEMIS Workshop*, pages 27–34, Leeds, UK, 2008.

[5] F. Bremond, M. Thonnat, and M. Zuniga. Video understanding framework for automatic behavior recognition. *Behavior Research Methods*, 3(38):416–426, 2006.

[6] C. Fernández, P. Baiget, X. Roca, and J. Gonzàlez. Natural language descriptions of human behavior from video sequences. In *In 30th Annual German Conference on Artificial Intelligence (KI-2007)*, Osnabrück, Germany, 2007.

[7] C. Fernández, P. Baiget, X. Roca, and J. Gonzàlez. Interpretation of complex situations in a cognitive surveillance framework. *Signal Processing: Image Communication Journal. Special issue on Semantic Analysis for Interactive Multimedia Services*, page To appear, 2008.

[8] J. H. Fernyhough, A. G. Cohn, and D. Hogg. Generation of semantic regions from image sequences. In *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume II*, pages 475–484, London, UK, 1996. Springer-Verlag.

[9] N. Ghanem, D. Dementhon, D. Doermann, and L. Davis. Representation and recognition of events in surveillance video using petri nets. In *In: Proceedings of Conference on Computer Vision and Pattern Recognition Workshops CVPRW*, page 2004, 2004.

[10] J. Gonzàlez and J.J. Villanueva. Understanding dynamic scenes based on human sequence evaluation. *Image and Vision Computing*, doi: 10.1016/j.imavis.2008.02.004, 2008.

[11] W. Hu, X. Xiao, Z. Fu, and D. Xie. A system for learning statistical motion patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1450–1464, 2006. Fellow-Tieniu Tan and Member-Steve Maybank.

[12] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC '95: Proceedings of the 6th British conference on Machine vision (Vol. 2)*, pages 583–592, Surrey, UK, UK, 1995. BMVA Press.

[13] D. Makris and T. Ellis. Path detection in video surveillance. *Image and Vision Computing*, 20:895–903, 2002.

[14] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems Man and Cybernetics–Part B*, 35(3):397–408, June 2005.

[15] C. Piciarelli and G. L. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recogn. Lett.*, 27(15):1835–1842, 2006.

[16] F. Porikli. Clustering variable length sequences by eigenvector decomposition using hmm. In *In Lecture Notes in Computer Science*, page 352. Springer-Verlag, 2004.

[17] C. Stauffer, W. Eric, and L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.