Recognition of scenarios describing human activities

François Brémond and Monique Thonnat
I.N.R.I.A., 2004 rte des Lucioles - B.P. 93
06902 Sophia Antipolis Cedex, France
Tel: (33) 04.93.65.76.59 - Fax: (33) 04.93.65.79.39
Francois.Bremond@sophia.inria.fr

Keywords: Scenario recognition, scene interpretation, knowledge-based system, real-word application.

Abstract

This paper describes a system for image sequence understanding. More precisely we address the issue of scenario recognition for scene interpretation systems. We start by presenting main issues in describing human activities and by presenting works related to these problems. Then based on a model, we represent a scenario describing human activities as a combination of sub-scenarios and properties on the mobile objects involved in the scenario. Depending on the type of combination (temporal or non temporal) we use an abductive diagnosis or an automaton as methods to recognize scenarios. Finally we propose a mechanism to help with the automatic generation of scenario recognition methods based on scenario descriptions.

1 Introduction

In this paper we address the issue of the representation of scenarios describing human activities for scene interpretation systems. The class of applications we are interested in, is the automatic surveillance of indoor and outdoor partially structured scenes with a fixed monocular color camera. Given image sequences describing a scene, an interpretation system has to recognize scenarios relative to the behaviors of mobile objects. In our case the mobile objects correspond either to humans or to vehicles and the scenarios describe some human activities.

We have developed an interpretation system which is composed of three modules. First, from image sequences the low-level image processing module subtracts the current image with the background image and detects moving regions. Then the tracking module tracks the detected regions and computes their trajectory. Finally the scenario recognition module identifies the tracked moving regions as mobile objects and interprets the scenarios that are relative to their behaviors.

In this paper we focus on the third module. We start by presenting related works on scenario representation [1], [4]. Then we explain how we represent scenarios based on sub-scenarios and we describe methods for the scenario recognition. We conclude by presenting

[0]In Proc. of the International Workshop on Dynamic Scene Recognition from Sensor Data, in Toulouse (France), June 1997

a mechanism to help with the automatic generation of scenario recognition methods. This issue is ambitious but essential as soon as we do not want to develop a totally new scenario recognition module while changing applications.

2 Related Works

The issue of describing human activities is at the border of the natural language domain and computer vision domain. First, as human activities are described based on natural language, two kinds of problems have to be solved :

- to define the referential aspect of the human activity (e.g. to decide whether the activity is performed relatively to a mobile object or to a static object belonging to context (also called the scene environment) [2]),
- to determine a set of basic elements to represent the numerous natural language terms used to describe human activities, even when activities are limited to a specific application domain (e.g. football games, car-park video-surveillance).

Second, as human activities match the real-world and are perceived through an image processing module, three other kinds of problems have to be tackled :

- to link the numerical properties of mobile objects with the symbolic descriptions of human activities (e.g. to select properties computed with image processing tasks in order to describe the action *"to sit down"*),
- to handle the incompleteness and uncertainty of mobile objects properties (e.g. to determine the reliability of the information on a man partially occluded by a car),
- to establish the duration of activities that depends on context (e.g. to estimate the time it takes to park a car in a cluttered scene).

Despite there is not a global solution for the issue of describing human activities, many works tackled some of these problems. For example, A. Galton in [6] generates complex descriptions of human actions based on a set of generic basic spatio-temporal propositions. B. Neumann in [8] states that symbolic descriptions have to be linked with properties defined at the image level and he describes car scenarios based on a pyramidal hierarchy of motion verbs with elementary motion verbs at the base of the pyramid (corresponding to simple events) and complex ones at its top (corresponding to scenarios).

R. Howarth in [7] emphasizes the difference between deictic, intrinsic or extrinsic referential. In human activities, the observer is usually the camera and mobile objects can have two roles : *source* or *target*. The source object is a mobile object that performs the action associated with the scenario. The target object is the reference of the action. The target object can be either a mobile object or a static object belonging to the scene context. For example, in the activity "*the man* goes round *the car*", *"the man"* is the source object and *"the car"* is the target object. For R. Howarth, humans are used to describe activities in a deictic referential (i.e. the observer is the source object) and so he translates mobile object properties in the deictic referential to represent car scenarios.

In the Perception project [5], they predefine plans of scenarios to analyze human activities in video-surveillance applications. Instantiated scenario plans are represented by Petri Nets. When the image processing module detects an event, they classify it and try to link it to an already existing scenario through links corresponding to structured and logical dependencies. For example, if the mobile object associated with the event is involved in the scenario (the event has a structured dependency) and if the event corresponds to the next stage of the scenario plan (the event has a logical dependency), then the event is linked to the scenario. They also model incomplete scenarios with partially instantiated scenario plans. Therefore they link a numerical information, the event, to a symbolic description, the scenario.

In this article we try to give a formalism to tackle most of the problems involved in the description of human activities.

3 Model of scenarios describing human activities

In our interpretation system the image processing module detects the moving regions and computes on them several measures (e.g. their size, location). Then the tracking module tracks the detected regions. Because of detection errors, a **moving region** can either correspond to a part of a mobile object (e.g. the arm of a person), to one mobile object (e.g. a person) or to a group of mobile objects (e.g. a crowd). Then the scenario recognition module generates hypothesis to consider the tracked moving regions as mobile objects composed of one or more regions. Finally the scenario recognition computes the properties of mobile objects and analyzes the scenarios relative to the behavior of mobile objects based on their properties. Up to now we are mainly using eight **mobile object properties** : height, width, speed, motion direction, current location, trajectory (set of previous locations), the distance to a reference object and the number of moving regions that compose the mobile object. The goal of the scenario recognition module is to recognize scenarios based on the properties computed on mobile objects.

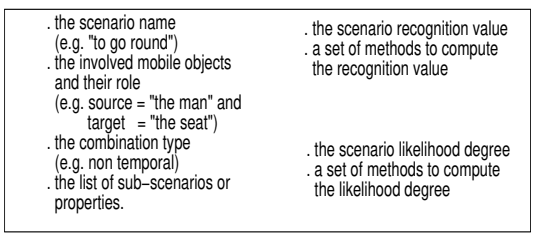


Figure 1: The model of scenario.

Thus we propose to recursively recognize a **scenario** relative to mobile objects based on the mobile object properties. At level 0, a scenario is recognized directly from the associate mobile object properties. At level n, a scenario is recognized through a combination of subscenarios recognized at level *n-1*. There are two types of combination : *non temporal* and *temporal*. First, a scenario can correspond to a non temporal constraint on a set of subscenarios and on mobile object properties. Second, a scenario can correspond to a temporal sequence of sub-scenarios. In the case of a temporal combination, the recursive definition

of scenarios allows us to easily change the scenario duration. For example, a scenario at an upper level can always be defined to add new sub-scenarios. For the scenario *"to park a car"*, we can define at an upper level a scenario to describe several attempts to park the car before succeeding.

To describe a scenario we use a model, shown on figure 1, composed of eight parts : the scenario name, the involved mobile objects, the combination type (*non temporal* or *temporal*), the list of sub-scenarios, the scenario recognition value, a set of methods to compute the recognition value, the scenario likelihood degree and a set of methods to compute the likelihood degree. For example, the scenario *"the man goes round the seat"* represents a non temporal combination of two properties : the distance between the man and the seat and the location of the man relative to the seat (the location value is one of the four cardinal points). If the distance is constant and if the relative location changes to another valid value, then the scenario is recognized.

We have used Sablayrolles's work to establish the scenario model. In [9] P. Sablayrolles gives several approaches to classify French motion verbs. For the viewpoint of verb representation, he divides 210 verbs into ten classes based on the spatial references of verbs. For example, one class gathers all the verbs describing actions starting out the spatial zone of reference and finishing inside the reference zone. The verb *"going into"* is a prototype of this class. In our scenario model through the roles and trajectory of mobile objects we can distinguish the ten proposed verb classes. As said before, a target object can be either a mobile object or a static object belonging to context (e.g. *"a seat"* or *"an exit zone"*). This classification enables also to determine the basic mobile object properties needed to recognize the class of a scenario. However specific properties may need to be added to the interpretation system when specific scenarios have to be analyzed. For this reason, we have developed a system that can be extended with new mobile object properties. Thus we expect that most of human activities involved in video-surveillance applications can be described by the proposed scenario model.

4 Scenario recognition methods

In this section we propose two main methods to recognize scenarios depending on the type of the combination of sub-scenarios.

4.1 Scenario representing a non temporal combination

If the scenario represents a non temporal constraint, then the scenario recognition value quantifies the constraint verification. In this case, the main point of the scenario recognition is to combine the values associated to sub-scenarios and properties in order to compute the scenario recognition value. However the hard point of the method is to handle the uncertainty of these values, mainly due to the inaccuracy of mobile object properties or to detection errors. To tackle this problem we compute the likelihood degree of a scenario through a diagnosis stage. Based on an abductive reasoning we diagnose whether the recognition value is really related to the scenario or whether the value is caused by other disorders. First, we determine all possible disorders that may interfere with the current scenario. Then we compute the symptoms which are related to the possible disorders. These symptoms can be seen as clues that indicate the intensity of the disorders. For

example, in the previous scenario *"the man goes round the seat"*, the symptoms correspond to detection errors due to several potential disorders like : the occlusion (either dynamic or static) of the mobile object at the level of its bottom, or the mix of the mobile object with another object while being tracked. So, using the observed symptoms we determine the intensity of the related disorders in the framework of fuzzy sets as described in [3]. Then we combine the intensity of the related disorders to quantify the diagnosis. Finally, according to the diagnosis result the likelihood degree of the scenario is increased or decreased. If the degree is too low the scenario is not taken into account and if it is high enough the scenario is said to be recognized.

4.2 Scenario representing a temporal combination

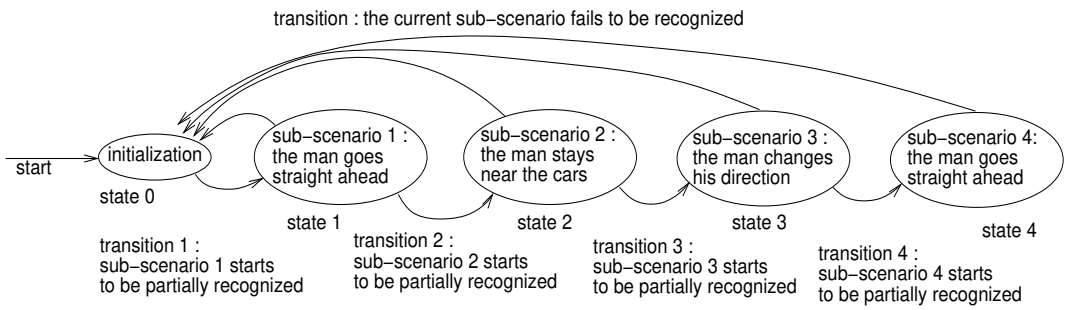


Figure 2: The figure shows the automaton of the scenario *"the man prowls around cars"*. There are five states of recognition : state 0 is used for the initialization and the other states correspond to the recognition of the sub-scenarios.

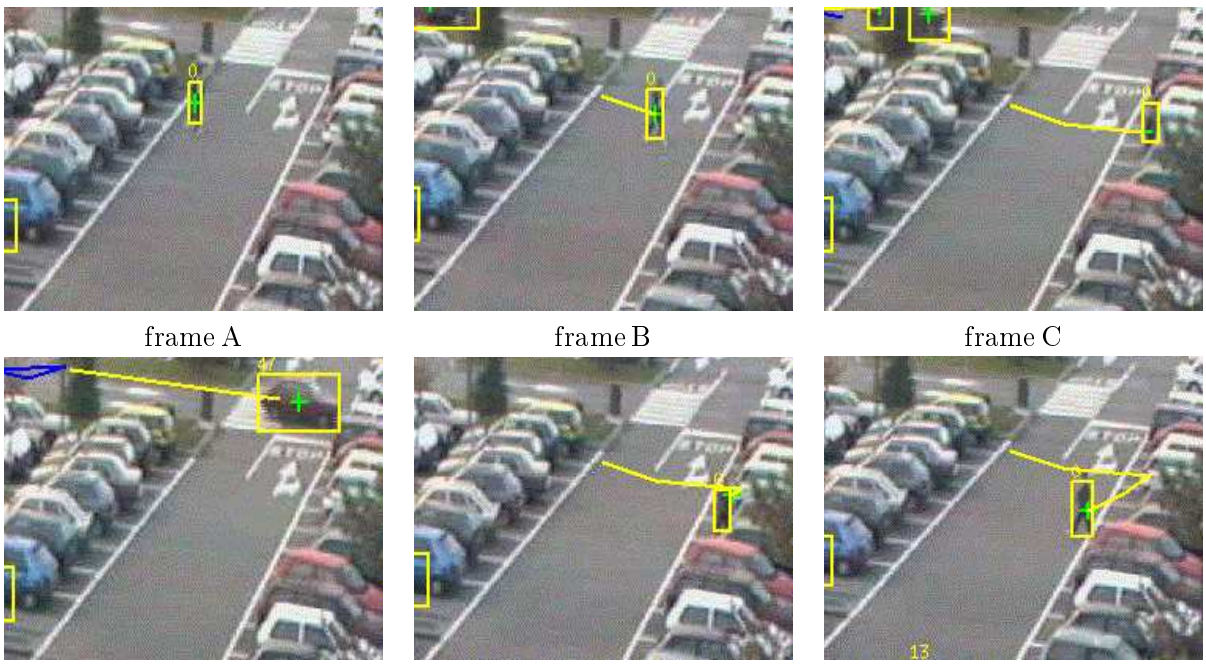
If the scenario represents a temporal sequence of sub-scenarios, then it is recognized through an automaton, the states of which correspond to the sub-scenarios. The scenario recognition value is the current state of recognition. The likelihood degree and the automaton transitions are computed through the likelihood degree of its sub-scenarios. The scenario is recognized when all its sub-scenarios are consecutively recognized and if its likelihood degree is high enough. For example, we have built a scenario that diagnoses whether or not a man is prowling around cars. The recognition automaton is described on figure 2. The scenario is composed of four sub-scenarios : *"the man goes straight ahead"*, *"the man stays near the cars"*, *"the man changes his direction"* and *"the man goes straight ahead"*. The scenario is recognized when the four sub-scenarios are consecutively recognized and if the likelihood degree is high enough. Thus likelihood degrees are propagated from bottom to top through the whole scenario recognition module. When a recognized scenario is interesting for a given application, then an alarm is triggered.

We do not use spatial or temporal logic as it is the case in several interpretation systems, because these logics cannot be used in all surveillance applications. In an idealistic situation no diagnosis stage is necessary and properties computed on mobile objects are true or false at one location and at one time-point. For example, this idealistic situation occurs when using an optical barrier as sensor. The instant when the optical barrier detects an event is a symbolic information that can be manipulated in the framework of a temporal logic. However this is not the case when using a camera and the diagnosis stage is then

often necessary. For example, to recognize the scenario *"the man has not moved for more than 3 minutes"* we need at least a diagnosis stage to establish when the scenario starts. Thus the manipulation of this information is uncertain and less informative. Therefore to represent the temporal aspect of scenarios, we just use the notion of a temporal sequence. The utilization of automaton is then sufficient to recognize a scenario corresponding to a temporal combination.

So, the diagnosis and the automaton are two methods that enables us to recognize all the scenarios described by the proposed scenario model.

5 Results



frame A frame B frame C

frame D frame E frame F

Figure 3: This image sequence shows a man prowling around cars.

One of our target applications is the video-surveillance of car-parks. As an example to validate the scenario recognition module, this section describes the recognition of the scenario *"the man prowls round cars"* from the image sequence of figure 3. These images have been taken in the framework of the European Esprit PASSWORDS project. The man starts to be detected on frame A. As he maintains his direction on frame B, the sub-scenario 1 begins to be recognized and the automaton goes from state 0 to state 1. On frame C, the man keeps going straight ahead and the likelihood of the scenario increases. On frame D, the man disappears behind cars, so we considere that is staying near the cars and the automaton goes to state 2 (sub-scenario 2 is recognized). On frame E, the man reappears after a while and changes his direction, so the automaton goes to state 3 (sub-scenario 3 is recognized). On frame F, as the man has kept his direction then the automaton goes to state 4 (the final state). All sub-scenarios have been consecutively and correctly

recognized, so the likelihood degree of the scenario *"the man prowls round cars"* is high and then this scenario is said to be recognized. On next frames, the man changes again his direction, so the automaton goes back to state 0 (the initial state) and the scenario will be able to be recognized a second time.

This example shows how to recognize a scenario composed of a temporal combination of sub-scenarios. The hard point in this recognition is to recognize sub-scenarios with a high likelihood degree and at the good time.

6 Discussion

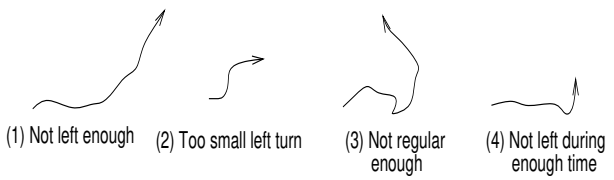


Figure 4: This figure shows four trajectories of a man that may have turned left depending on context.

The previous section describes scenario recognition methods and explains how they work. But our goal is to use the presented interpretation system in different video-surveillance applications. Thus a main point is to generate automatically the scenario recognition method (the abductive diagnosis or the recognition automaton) for a particular scenario given its description. The mechanism to generate a recognition automaton is classical. First, given the sequence of sub-scenarios we have to build the states and transitions of the automaton. Second, we have to decide when a sub-scenario (state i) is sufficiently recognized to try to recognize the next sub-scenario (state i+1). Thus this problem consists in fixing thresholds of recognition degree. However the problem of generating abductive diagnosis does not have a simple solution because of two main difficulties. First, the diagnosis stage of a scenario consists in establishing the likelihood degree of the scenario, knowing that several disorders can also explain similar situations. For example, consider the two scenarios : scenario 1 *"the man sits down"* and scenario 2 *"the man squats down"*. If scenario 1 is the only possible scenario, then to recognize it we just need to check that the size of the man shrinks down. If scenario 2 has also to be considered, then the scenario recognition method of scenario 1 has to change. So to define a scenario recognition method we need to know the description of the current scenario and the description of all disorders that may interfere with the scenario. Thus no library of scenario recognition methods can be predefined, because of the need to know before all the descriptions of the scenarios involved in the scene. The second difficulty in generating abductive diagnosis is that sometimes scenarios depend on the context of the scene. As a scenario recognition method is built to match the meaning associated to the scenario, then when the scene changes the scenario meaning also may change. For example, figure 4 shows different trajectories that can correspond to the scenario *"the man turns left"* depending on the scene context. Thus a predefined library of scenario descriptions may be no more valid when changing context. Because of these two difficulties, we think that the best mechanism to

generate diagnosis is to use predefined and generic scenario descriptions. These descriptions have to be seen as prototypes that not depend on context. So the mechanism to generate diagnosis consists first in tuning the descriptions to match the current context. Then diagnosis can be generated based on the whole set of tuned descriptions. However this mechanism requires an important experience in the domain of scene interpretation. For this reason we do not have for the moment generated automatically scenario recognition methods based on predefined scenario descriptions.

7 Conclusion

This paper explains how scenarios describing human activities are recognized in our system of scene interpretation. A scenario is recursively defined as a combination of sub-scenarios and properties on the mobile objects involved in the scenario. Then we propose two main methods to recognize a scenario depending on the type of the combination. As our goal is to develop real-world applications, we face another problem : when changing the scene, scenarios may also change. For this reason, we propose a mechanism to generate automatically scenario recognition methods. The mechanism consists in using a library of predefined scenario descriptions and in adapting them automatically to a new scene.

Therefore in our future works we first plan to develop several video-surveillance applications using this scenario recognition module. Then we will try to implement the mechanism to generate automatically scenario recognition methods from generic scenario descriptions.

References

[1] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *Workshop on Applications of Computer Vision*, December 1996.

[2] F. Brémond and M. Thonnat. A context representation for surveillance systems. In *European Conference on Computer Vision, workshop on Conceptual Descriptions from Images*, Cambridge, April 1996.

[3] F. Brémond and M. Thonnat. Analysis of human activities described by image sequences. In *the 10th international FLAIRS Conference*, Florida, May 1997.

[4] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1-2):431–459, 1995.

[5] C. Castel, L. Chaudron, and C. tessier. What is going on? A high level interpretation of sequences of images. In *ECCV'96 workshop on Conceptual Descriptions from Images*, University of Cambridge, April 1996.

[6] A. Galton. Towards an integrated logic of space, time and motion. In *IJCAI, Chambery, France*, page 1550. International Joint Conference on Artificial Intelligence, August 1993.

[7] R. Howarth. *Spatial representation, reasoning and control for a surveillance system.* PhD thesis, Queen Mary and Westfield College, July 1994.

[8] B. Neumann. *Semantic structures : advances in natural language processing*, chapter 5, pages 167–206. David L. Waltz, 1989.

[9] Pierre Sablayrolles. *Sémantique formelle de l'expression du mouvement. De la sémantique lexicale au calcul de la structure du discours en français.* PhD thesis, Thèse IRIT - Université Paul Sabatier Toulouse, 1995.