ANALYSIS OF HUMAN ACTIVITIES DESCRIBED BY IMAGE SEQUENCES

François Brémond and Monique Thonnat
I.N.R.I.A., 2004 rte des Lucioles - B.P. 93
06902 Sophia Antipolis Cedex, France
Francois.Bremond@sophia.inria.fr

Abstract

This paper describes a system for image sequence understanding. More precisely we address the issue of scenario recognition for scene interpretation systems. After presenting related works on scenario recognition, we first establish different requirements to describe human activities involved in video-surveillance applications. Then based on a model, we define a scenario as a combination of sub-scenarios. Depending on the type of combination (temporal or non temporal) we use an abductive diagnosis or an automaton to recognize the scenario. Then we explain through an example how to use the scenario representation. Finally we propose several mechanisms to obtain a real-time scenario recognition process.

1 FRAMEWORK

In this paper we address the issue of knowledge representation and real-time control 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 [Castel *et al.*, 1996]. In our case the mobile objects correspond either to humans or to vehicles and scenarios describe human activities.

We have developed an interpretation system which is composed of three modules. First, from image sequences the low-level image processing module detects moving regions. Then the tracking module tracks the detected regions. 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 show how the knowledge of the scenario recognition is represented and how a real-time control of the recognition process is developed.

Knowledge representation is an important issue for the scenario recognition module because of the characteristics of target applications. As these applications are dealing with real-world images, the module has to manipulate uncertain and incomplete data [Brémond and Thonnat, 1996b], use contextual information (e.g. also called the scene environment) [Brémond and Thonnat, 1996a] and choose the temporal granularity of the scenarios to recognize [Nagel, 1991]. Moreover scenarios are described by human operators and are expressed in natural language. So the module has to manipulate scenarios involving several objects, to define the references used to describe scenarios, and to use spatio-temporal constraints to represent them.

The second issue is the recognition control in order to obtain a real-time system. As every image needs an analysis, the module has to adapt its process to the available time [Buxton and Gong, 1995].

This paper presents a scenario recognition module that partly answers these two issues. First we explain how we represent the knowledge of the scenario recognition. Then we show some mechanisms that make possible the real-time control of the module. We conclude by giving an example of its utilization.

2 SCENARIO REPRESENTATION

2.1 Related Works

The issue of representing scenarios relative to mobile objects has been especially tackled in the case of rigid objects (e.g. vehicles or robots). Rigid objects have often the advantage of making sharp and clear move-

0

ments. Thus scenarios can be defined as a sequence of events, each event corresponding to a change at one time-point in the scenario. There are different methods to split a scenario into events. For example, in an application of car surveillance [Nagel, 1991] has built a pyramidal hierarchy of motion verbs with elementary verbs at the base of the pyramid (corresponding to events) and complex ones at its top (corresponding to scenarios). [Dousson *et al.*, 1993] have recursively defined scenarios as sequences of sub-scenarios from a set of basic events. Thanks to this formalism they have developed an application based on temporal constraint propagation for the surveillance of a mobile robot.

The representation of scenarios involving non rigid objects (e.g. humans) is more difficult because in most situations events cannot be accurately defined. However there are several methods to overcome this problem. For example [Kuniyoshi and Inoue, 1993] have limited scenario representation to structured scenes and they have managed to recognize the action of a hand assembling blocks on a table. Another method is to recognize particular scenarios. For example, [Bobick and Davis, 1996] have defined different models of actions such as *"the man is sitting"*. But these models have to be explicitly defined for every viewing angle of the camera.

In this article we try to give a formalism to represent most of human activities involved in video-surveillance applications.

2.2 Describing Human Activities

In human activities, mobile objects can have two roles : *source* or *target*. The source object is a mobile object that realizes 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 context of the scene (or scene environment). Depending on the combination of roles we have defined six types of scenarios :

- Scenarios with a target object belonging to **context** and **near** the source object. For example, in the scenario "*the man* goes round *the seat*", *"the man"* is the source object and *"the seat"* is the target object.
- Scenarios with a target object belonging to **context** and **far** from the source object (e.g. "*the man* goes towards *the door*").
- Scenarios with a **mobile** target object **near** the source object (e.g. "*the man* touches *the other man*").
- Scenarios with a **mobile** target object **far** from the source object (e.g. "*the man* follows *the other man*").
- Scenarios with **just a source** object corresponding to **one** real object (e.g. "*the man* walks then comes back").
- Scenarios with **just a source** object corresponding to **a group** of real objects (e.g. "*the group of men* scatters then gathers").

These scenario types allow to describe most of human activities.

In our interpretation system the image processing module detects the moving regions and computes several measures (e.g. their size, location). Then the tracking module tracks the detected regions. 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 (e.g. *"height"*) or the evolution of their properties (e.g. *"the height increases"*). 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 the mobile objects.

2.3 Scenario Model

. the scenario name
(e.g. "to go round")
. the involved mobile objects
and their role
(e.g. source = "the man" and
target = "the seat")
. the combination type
(e.g. non temporal)
. the list of sub−scenarios or
properties.

. the scenario recognition value
. a set of methods to compute
the recognition value

. the scenario likelihood degree
. a set of methods to compute
the likelihood degree

Figure 1: The model of scenario.

We propose to recursively recognize a **scenario** relative to mobile objects based on 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 sub-scenarios recognized at level *n-1*. There are two types of combination : *non temporal* and *temporal*. First, a scenario can represent a non temporal constraint on a set of sub-scenarios and on mobile object properties. Second, a scenario can represent a temporal sequence of sub-scenarios. As shown on figure 1, a scenario is described by a model 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.

2.4 Scenario Representing a Non Temporal Combination

If the scenario represents a non temporal constraint, then the scenario recognition value quantifies the constraint verification. The main point of this scenario recognition is to combine the values associated to sub-scenarios and properties. However the hard point of the method is to handle the uncertainty of these values due to detection errors. To tackle this problem we use the likelihood degree of the scenario. The likelihood degree is diagnosed by combinating symptoms in the framework of fuzzy sets as described in [Brémond and Thonnat, 1996b]. For the preceding example the symptoms correspond to detection errors due to several potential disorders : is the mobile object occluded (either dynamic or static occlusion) at the level of its bottom? has the mobile object been mixed with another object while being tracked?... The diagnosis stage is divided into four steps. First, we compute the membership degree of all possible symptoms s_j to the set of the observed symptoms, noted $\mathcal{S}_{\mathcal{OBS}}$. This membership degree, noted $\mu_{\mathcal{S}_{\mathcal{OBS}}}(s_j)$, measures the current intensity of the symptom s_j. Second, we compute the membership degree of all possible disorders d_i to the set of the solution disorders (noted $\mathcal{D}_{\mathcal{SOL}}$) with the following formula :

$$\mu_{\mathcal{D}_{\mathcal{SOL}}}(d_i) = max_j[\,min(\mu_{\mathcal{S}(d_i)}(s_j),\, \mu_{\mathcal{S}_{\mathcal{OBS}}}(s_j))\,]$$

The coefficient $\mu_{\mathcal{S}(d_i)}(s_j)$ quantifies the implication : *the disorder d_i causes the symptom s_j*. This coefficient belongs to the a priori knowledge of the application. The membership degree $\mu_{\mathcal{D}_{\mathcal{SOL}}}(d_i)$ measures the current intensity of the disorder d_i. Third, we compute the average intensity of d_i at time t :

$$\text{intensity}_{d_i}(t) = [\mu_{\mathcal{D}_{\mathcal{SOL}}}(d_i) + \text{intensity}_{d_i}(t-1)]\ /\ 2$$

Fourth, we compute the degree of the average likelihood at time t thanks to the current likelihood degree :

$$\text{crt-likelihood} = 1 - max_i[\min(\beta_i,\ \text{intensity}_{d_i}(t))\,]$$
$$\text{likelihood(t)} = [\text{crt-likelihood} + \text{likelihood(t-1)}\ ]\ /\ 2$$

β_i is a coefficient that belongs to the a priori knowledge. It measures the importance of the disorder d_i for the scenario. Thus 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.

2.5 Scenario Representing a Temporal Combination

If the scenario represents a temporal sequence, then it is recognized through an automaton, the states of which represent 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 looking for a seat. The recognition automaton is described on figure 2. The scenario is composed of two sub-scenarios : *"the man goes round the seat"* and *"the man sits down"*. The scenario is recognized when the two 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.

This scenario representation has several advantages. First, the recursive definition of the recognition process facilitates the description of scenarios that last as long as needed. Second, the degree of likelihood enables to obtain a more robust scenario recognition. Third, the benefit of having six types of scenarios is to be able to represent most of the scenarios describing human activities as it is done in natural language. Thus this knowledge representation enables us to describe and to recognize scenarios.

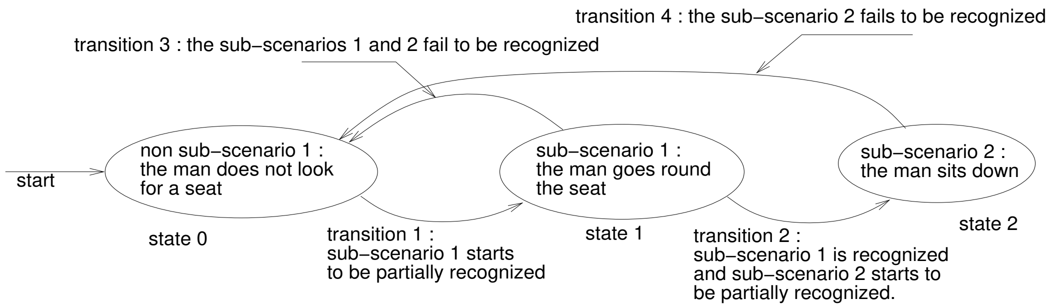

Figure 2: The figure shows the automaton of the scenario *"the man looks for a seat"*. There are three states of recognition : state 0 is used when sub-scenario 1 has not yet started to be recognized, state 1 and 2 correspond to sub-scenario 1 and 2.

3 REAL-TIME CONTROL

Due to application characteristics (video-surveillance) we need to define the mechanisms that analyze a sequence in real-time. For this we use two modes : data driven and goal driven mode. The **data-driven** mode (or bottom-up mode) realizes a quick processing on the whole scene allowing the system to detect and to track all mobile objects and therefore to give a minimal analysis of the scene. The **goal-driven** mode (or top-down mode) selects the most interesting mobile objects and performs sophisticated scenario recognitions on them. The first mode is automatic and is performed on every image whereas the second one is used only if the system has time.

To implement these two modes we have defined four operations (ordered from a to d). *Operation a* computes the degree of interest for each mobile object according to its certainty degree and to the likelihood degree of its associate scenarios. *Operation b* gives the threshold of interest, from which a mobile object is considered as interesting. *Operation c* selects the suitable scenario associated to a mobile object according to the object properties. *Operation d* computes the time frequency indicating when the goal-driven mode is used.

Operations a and b are used to compute the list of mobile objects of interest and operation c indicates which recognition processes to perform on these mobile objects. These three operations give a mechanism to implement the goal-driven mode. As the data-driven mode is controlled by the data flow it does not need additional operations. Operation d establishes the proportion between the two modes. If for example the time frequency is small then only few mobile objects will have sophisticated scenario recognitions, but these recognitions will be precise. So these four operations allow us to implement a real-time control of the scenario recognition.

4 RESULTS

One of our target applications is the video-surveillance of metro stations. As a first example to validate the scenario recognition module we try to recognize scenario 1 *"the man looks for a seat"*, described on figure 2, which is composed of sub-scenarios 1 and 2. Figure 3 shows three frames used for the recognition of this scenario. They are part of an image sequence taken by SRWT in a metro station in Charleroi in the framework of the esprit PASSWORDS project [Chleq and Thonnat, 1996]. These scenarios involve two objects : a mobile source object (*"the man"*) and a target object that belongs to context and that is near the source object (*"the seat"*). The first step starts with the recognition of sub-scenario 1. So the scenario recognition module computes the object properties used for this analysis : the distance between the man and the seat and the location of the man relative to the seat. From frame 1 to frame 3 the distance between the man and the seat is constant and the relative location change from North (on frame 1) to West (on frame 2) and then to South (on frame 3). Thus sub-scenario 1 is recognized at frame 2. Because of this success the module considers the man as a mobile object of interest and performs on it a process at a higher level of recognition. So the second step is to recognize scenario 1. As the recognition automaton of scenario 1 is on state 1, the recognition process tries to recognize sub-scenario 2. This sub-scenario is characterized by two properties : the location of the man has to be constant and near the seat, and its size has to shrink. As these two properties are not verified, then sub-scenario 2 and hence scenario 1 are not recognized. So the recognition automaton of scenario 1 goes from state 1 to state 0.

Thus the result of the scenario recognition at the end of the sequence is that sub-scenario 1 *"the man goes round the seat"* is recognized and that scenario 1 *"the man looks for a seat"* is not recognized.

Figure 3: From frame 1 to frame 3, a man is going round a seat. The direction of his movement is represented by an arrow.

5 CONCLUSION

This paper proposes a knowledge representation and a real-time control to recognize scenarios describing human activities. A scenarios is recursively defined as a combination of sub-scenarios and there are two ways to recognize a scenario depending on the type of the combination. For a non temporal combination we use an abductive diagnosis and for a temporal combination we use a recognition automaton. Thanks to this formalism, our scenario recognition module is able to recognize the scenario *"the man goes round a seat"* as it was required. We think that this formalism can describe most of scenarios involved in video-surveillance applications. However, a main drawback is that all the scenario recognition methods (diagnosis and automatons) belong to the a priori knowledge and have to be explicitly described for each application.

Therefore in our future works we will first try to use the scenario representation in other video-surveillance applications to determine how generic is the formalism. Then we will try to define generic libraries of scenario recognition methods and to use them in different applications. Another goal is to generate automatically scenario recognition methods from descriptions in natural language. This goal is ambitious because scenarios are described based on mobile object properties and these properties are not always directly linked to scenarios.

References

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

[Brémond and Thonnat, 1996a] 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.

[Brémond and Thonnat, 1996b] F. Brémond and M. Thonnat. Interprétation de séquences d'images et incertitude. In *Proc. of Logique Floue et Applications LFA*, Nancy, 1996.

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

[Castel *et al.*, 1996] 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.

[Chleq and Thonnat, 1996] N. Chleq and M. Thonnat. Realtime image sequence interpretation for surveillance applications. In *IEEE International Conference on Image Processing, ICIP*, pages 801–804, Lausanne, Switzerland, September 1996.

[Dousson *et al.*, 1993] C. Dousson, P. Gabarit, and M. Ghallab. Situation recognition representation and algorithms. In *IJCAI-93, chambéry*, volume 1, pages 166–172, August 1993.

[Kuniyoshi and Inoue, 1993] Y. Kuniyoshi and H. Inoue. Qualitative Recognition of Ongoing Human Action Sequences. In *Int'l Joint Conference on Artificial Intelligence*, volume 2, pages 1600–1609, Chambéry, August 1993.

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

[Nagel, 1991] H. Nagel. The representation of situations and their recognition from image sequences. In *RFIA Lyon Villeurbanne*, pages 1221–1229, November 1991.