Scenario Recognition in Airborne Video Imagery

Francois Brémond Gérard Medioni
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA 90089-0273
{bremond,medioni}@iris.usc.edu

Abstract

We describe here 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 work 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 check whether some specific constraints are verified, or we use automata as methods to recognize scenarios. We then illustrate how the scenario recognition module works through an example of utilization. Finally, we underline the major problems we have to face with the development of scenario recognition methods.

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

1 Introduction

We address here the issues of the representation of scenarios describing human activities for scene interpretation systems. The class of applications we are interested in, is the automatic surveillance and monitoring of outdoor scenes observed by a camera located on an airborne platform. 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 human activities.

We have developed an interpretation system which is composed of three modules as shown on figure 1. First, from image sequences the low-level image processing module (also called stabilization module) compensates the image flow induced by the motion of observation platform and detects moving regions based on optical flow techniques. 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. This third module uses two kinds of context (defined as a priori information on the scene environment and acquired before the processing of the scene) [5] : spatial context and mission context. The spatial context corresponds to a map of the scene. The mission context contains the specific methods to recognize the mission scenarios.

The two first modules are described in [8]. In this paper, we focus on the third module. We start by presenting related work on scenario representation [1], [11], [2], [6]. We then present our representation formalism : we explain how to represent scenarios based on mobile object properties, and describe the methods we have developed to recognize scenarios. We conclude by presenting an utilization example of our interpretation system, and by showing the advantages of using the proposed representation formalism. The experimentation in real-word applications is essential for us because this is the best way to see how efficient is a representation formalism.

2 Related Work

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 :

- The definition of the referential aspect of the human activity : how to decide whether the activity is performed relative to a mobile object, or to a static object belonging to the spatial context,
- The definition of a set of basic properties : 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, parking 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 :

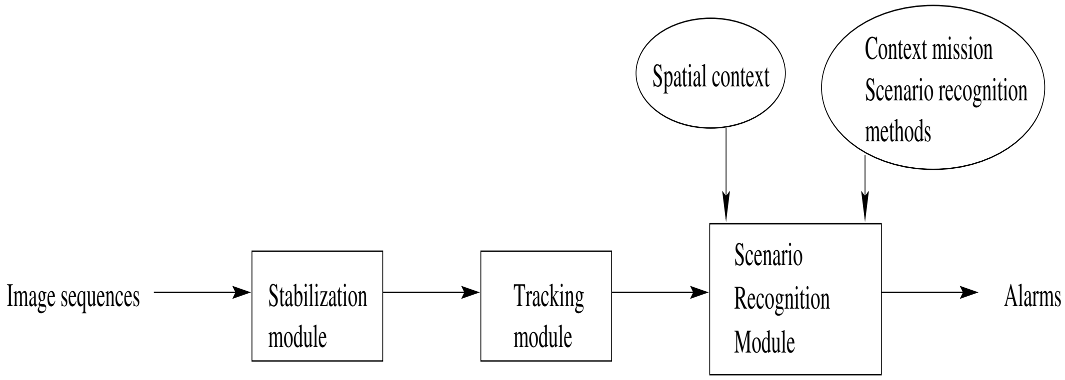


Figure 1: The interpretation system is composed of three modules.

- to bridge the gap between the numerical properties of mobile objects and the symbolic descriptions of human activities : to select properties computed with image processing tasks in order to describe an action such as *"to slow down"*,
- to handle the incompleteness and uncertainty of mobile object properties (e.g. to determine the reliability of the detection of a car partially occluded),
- to establish the duration of activities that depend on context and to segment these activities into elementary events (e.g. to estimate the time it takes to park a car in a cluttered scene and to decompose this activity into elementary events).

While there is no global solution for the issue of describing human activities, many works tackle some of these problems. For example, A. Galton in [9] generates complex descriptions of human actions based on a set of generic basic spatio-temporal propositions.

B. Neumann in [12] states that symbolic descriptions must 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 [10] emphasizes the difference between deictic, intrinsic or extrinsic referential. For this author, humans are used to describe activities in a deictic referential : the referential is the observer (i.e. the camera). For this reason he translates mobile object properties in the deictic referential to represent car scenarios.

In the Perception project [7], the authors 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 then try to associate it with a node that belongs to an already existing scenario. The edges in the Petri Net correspond to structured and logical dependencies between nodes. For example, if the mobile object associated with the event is involved in the scenario (the event is then said to have a structured dependency) and if the event corresponds to the next stage of the scenario plan (the event is said to have a logical dependency), then the event is linked to the scenario. They also model incomplete scenarios with partially instantiated scenario plans. Based on this representation they link a numerical information, *the event*, to a symbolic description, *the scenario*.

In this paper, we call scenario any activity related to humans. For the mobile objects involved in a scenario, we have defined two roles : *source* or *reference*. The source object is a mobile object that performs the action associated with the scenario. The reference object is the reference of the action. The reference object can be either a mobile object or a static object belonging to the scene context. For example, in the activity "*the car* goes toward *the checkpoint*", *"the car"* is the source object and *"the checkpoint"* is the reference object.

In our work, the goal is to give a representation formalism to tackle most of the problems (enumerated above) that are involved in the issue of scenario recognition.

3 Model of Scenarios Describing Human Activities

3.1 Mobile Object Properties

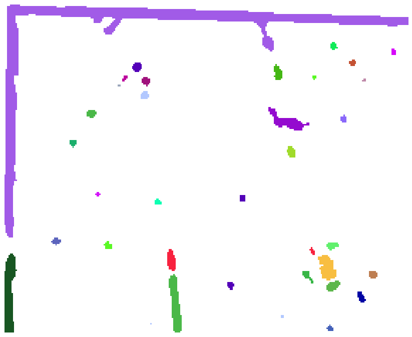
In our interpretation system, the image processing module detects the moving regions and computes on them several **measures**. Up to now we are mainly using eight measures : height, width, speed, motion direction, current location, trajectory (set of previous locations), the distance to a reference object and the number of moving regions that compose the mobile object. Then, the tracking module tracks the detected regions. Because of detection errors, as shown on figure 1, 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. We show on figure 2 the properties we use in our system. These properties have usually a numerical value and they are supposed to be instantaneous. They are generally computed on a short interval of time (typically 4 or 5 frames for video surveillance application) in order to avoid unstable values. The mobile object properties are intended to be generic (i.e. independent of applications) and re-usable for different application types. Thus we use these properties as basic elements to recognize scenarios.

3.2 Scenarios

For the scenario recognition module, we propose to recursively recognize a **scenario** based on the properties of the mobile objects involved in the scenario. 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 combinations : *non temporal* and *temporal*. First, a scenario can correspond to a non temporal constraint on a set of sub-scenarios 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 and the temporal segmentation. 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. Thus scenarios have a symbolic value, and we use them to recognize activities on long image sequences. They are intended to describe various human activities, and adapt the recognition to different types of applications.

To describe a scenario we use a model, shown on figure 4, 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 car goes toward the checkpoint"* represents a non temporal combination of three properties : the distance between the car and the checkpoint, the direction of the car and its speed. If the distance decreases, the direction is toward the checkpoint and the car slows down, then the scenario is said to be recognized.

input frame



moving regions detected

Figure 3: These two frames show the set of the moving regions detected by the first module. We can notice that an important number of detected regions correspond to noise.

We have used Sablayrolles's work to establish the scenario model. In [13], 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 of mobile objects and their trajectory we can distinguish the ten proposed verb classes. 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

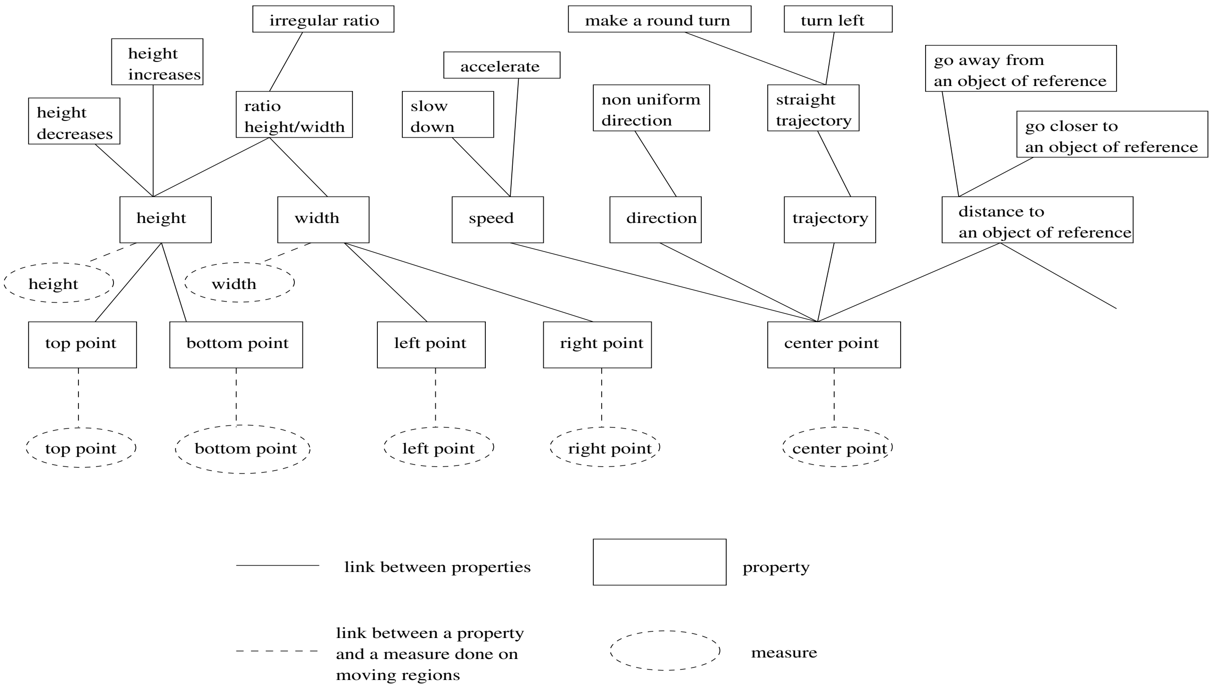


Figure 2: The set of mobile object properties.

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 of the associate sub-scenarios and properties in order to compute the scenario recognition value. However, the hard part 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 have defined a likelihood degree for scenarios. This is a numerical value indicating how reliable is the computation of the scenario recognition value.

We use two kinds of methods to compute the likelihood degree. Most of the time, and in particularly for long scenarios, we use the temporal coherency : during the processing, if the new scenario value is coherent with the old ones, we increase the likelihood degree of the scenario. For example, when computing the scenario "the car slows down" if we notice that the car speed decreases effectively at each new frame arrival, we incrementally increase the scenario likelihood degree. Second, to handle specific cases we sometimes compute the likelihood degree using possibilistic logic [3]. When a scenario is based on several mobile object measures, and when we cannot wait for the arrival of new information, we diagnose thanks to possibilistic logic, whether the scenario value computation is reliable or not at a given time point. For example, when we compute the scenario "the height decreases" we also have to diagnose whether the computation of the scenario value could have been influenced by the occlusion of the top of the mobile object. 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. So, using the observed symptoms we determine the intensity of the related disorders in the framework of fuzzy sets as described in [4]. 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.

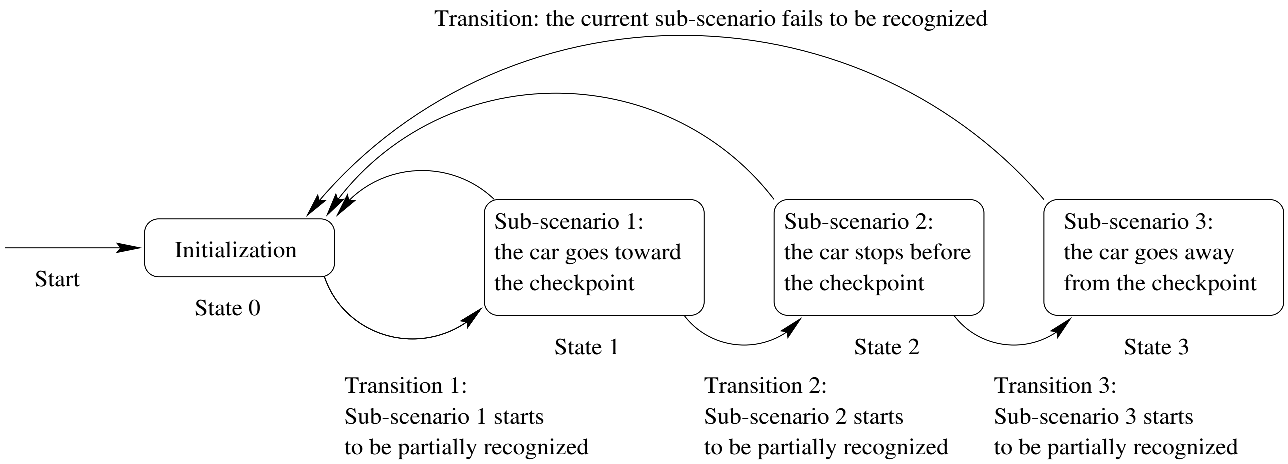


Figure 5: The figure shows the automaton of the scenario *"the car avoids the checkpoint"*. There are four states of recognition : state 0 is used for the initialization and the other states correspond to the recognition of the sub-scenarios.

the scenario name (e.g. "to pass the checkpoint").

the involved moving objects and their role
(e.g. source="the car" and reference="the checkpoint").

the combination type (e.g. 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 4: The model of scenario.

4.2 Scenario Representing a Temporal Combination

If the scenario represents a temporal sequence of sub-scenarios, then it is recognized through an automaton, the states whose 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 car is avoiding a checkpoint. The recognition automaton is described on figure 5. The scenario is composed of three sub-scenarios : *"the car goes toward the checkpoint"*, *"the car stops before the checkpoint"* and *"the car goes away from the checkpoint"*. The scenario is recognized when the three 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 car has stoped 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 of sub-scenarios.

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

5 Experimental Results

One of our target applications is video-surveillance of streams obtained by the Predator U.A.V. (a small airplane). We choose as instance of scenario the monitoring of checkpoints (i.e. road blocks). As an example to validate the scenario recognition module, this section describes two image sequences depicting car behaviors related to the monitoring of a checkpoint. The image sequences are shown on figure 6. They have been taken from an airborne platform in the framework of the V.S.A.M. project and acquired at 15Hz. On sequence A we can see *"a car passing through the checkpoint"* called scenario 1 (this is defined as a normal behavior) and on sequence B we can see *"a car avoiding the checkpoint"* called scenario 2 (this is defined as an abnormal behavior). In this paper we just describe sequence B. We have drawn two red polygons to delimit two contextual zones : the road and the checkpoint. On the frames the green rectangles correspond to the bounding boxes of the moving regions associated with the detection of the car and the yellow lines correspond to the car trajectory computed by the system.

On frame B19, the car starts to be detected. As the car is on the road the interpretation system considers it as a mobile of interest and we initiate the recognition of scenario 1 and scenario 2. These scenarios are predefined in the system.

On frame B39, the car slows down, maintains its direction toward the checkpoint and its distance to the checkpoint decreases so sub-scenario 2.1 starts to be recognized and the automaton of scenario 2 goes to state 1. This automaton is shown on figure 5.

On frame B107, the car stops before reaching the checkpoint (its speed is below a predefined threshold) so the automaton goes to state 2 (sub-scenario 2.2 is recognized).

On frame B208, the car backs out : its distance to the checkpoint increases, its direction is opposite to the checkpoint and the car is still on the road. Thus sub-scenario 2.3 is recognized and the automaton goes to state 3 (the final state).

All sub-scenarios have been consecutively and correctly recognized, so the likelihood degree of scenario 2 *"the car avoids the checkpoint"* is said to be recognized. Meanwhile scenario 1 *"the car passes through the checkpoint"* has failed to be recognized because the car did not stop at the checkpoint as required by sub-scenario 1.2. An abnormal behavior has been recognized with enough confidence, so the system triggers an alarm. This example shows how to recognize a scenario composed of a temporal combination of sub-scenarios. However, to obtain a good evaluation of the system, we need further tests on several image sequences. One of the most difficult issues in this type of recognition is the ability to define generic properties and scenarios that can be applied on a large number of sequences. What we have learned through our experimentations can be summarized in four points :

- The utilization of context is a key point. Contextual information is often more reliable, so we use it to compute the main properties : for example in sub-scenario 2.1 the distance to the checkpoint (defined as a contextual zone) is more reliable than the computation of the car speed.
- The recognition of scenarios requires flexibility. First, properties can be used in different ways to recognize scenarios. Second, we still want to recognize scenarios, even if we temporary lose the mobile object tracks.
- We usually do not need sophisticated scenario recognition methods, because accurate property values are often not available. For instance, we can compute accurately the car acceleration in one sequence such as sequence A (because here the car is sharply detected), but not necessarily in other sequences, such as sequence B, where we lose the car track many times.
- We still need accurate scenario recognition methods if we want to discriminate one given scenario from the others.

As we can see, point 2 and point 3 are in opposition with point 4. So the problem is to determine the good balance between these three points. At the moment we are more concerned by the development of scenario recognition methods, because these methods still require a tuning phase to obtain reliable results. For instance the recognition of the scenario *"the car stops"* means that the speed of the car is below a predefined threshold : because of detection instability a mobile object never really seems to stop. So we need to tune the speed threshold, depending on the average speed of scene objects.

Therefore, our current work consists in designing new techniques to automatically develop and tune these recognition methods. A second purpose is to make these methods generic.

6 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. We then 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 an important issue : a scenario recognition method has to be flexible enough to be able to recognize the given scenario in various image sequences but it has also to be accurate enough to just recognize the given scenario. To solve this problem we propose to use two temporal granularities of description (properties and scenarios) and to characterize scenarios through two values (the recognition value of the scenario and its likelihood degree). The major problem of this solution is that we have to adjust scenario recognition methods in order to obtain efficient results.

Therefore our current work consists in automating the generation of scenario recognition methods.

References

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

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *proc. of CVPR*, Puerto Rico, USA, 1997.

[3] F. Brémond and M. Thonnat. Interprétation de séquences d'images et incertitude. In *Proc. of the Rencontres sur la Logique Floue et ses Applications (LFA)*, Nancy, December 1996.

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

[5] F. Brémond and M. Thonnat. Issues of representing context illustrated by video-surveillance applications. *International Journal of Human-Computer Studies, Special Issue on Context*, 1998.

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

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

[8] I. Cohen and G Médioni. Detection and tracking of objects in airborne video imagery. Technical report, University of Southern California, 1998.

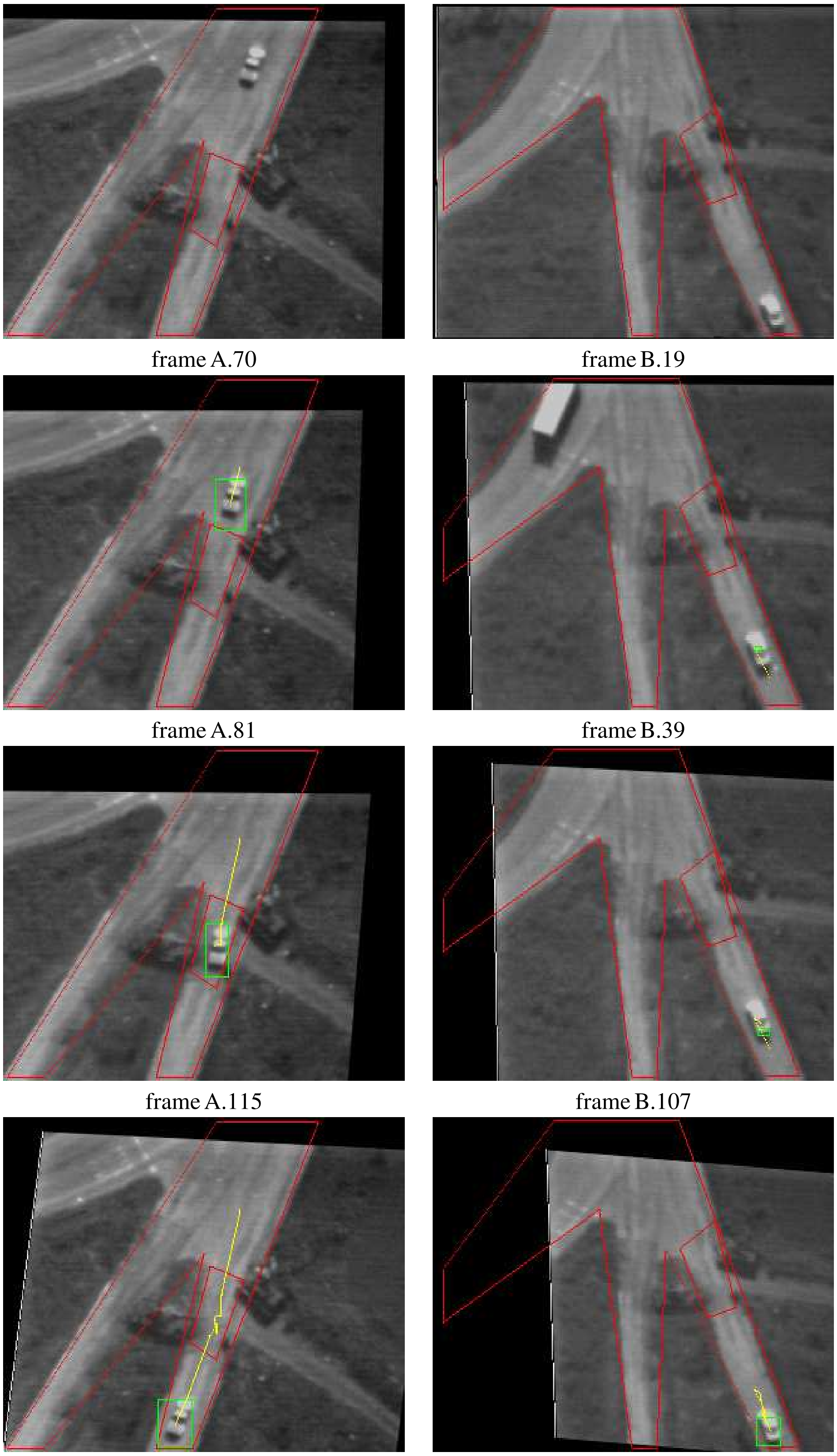
[9] A. Galton. Towards an integrated logic of space, time and motion. In *International Joint Conference on Artificial Intelligence (IJCAI), Chambery, France*, August 1993.

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

[11] R. Morris and D. Hogg. Statistical models of object interaction. In *proc. of the Int'l Conference on Computer Vision (ICCV) Workshop on Visual Surveillance*, Bombay, India, 1998.

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

[13] 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.



frame A.70 | frame B.19

frame A.81 | frame B.39

frame A.115 | frame B.107

frame A.163 | frame B.208

Figure 6: These two image sequences show car behaviors related to the monitoring of a checkpoint.