Perception of Action in Computer Vision

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

1 Issues in perception of action

1.1 What do we mean by action?

An action is any process performed by (at least) one mobile object. A mobile object can be, for example, a human, a group of humans, a vehicle or a robot. At a higher level, actions are used to understand the behaviors of mobile objects. An action is perceived through the computation of properties (and property evolutions) relative to the moving regions corresponding to the involved mobile objects. *"The man has stayed in the dangerous zone for more than three minutes"*, *"the man size shrinks"* and *"the group of men scatters"* are examples of properties (or property evolutions). The property relative to an action is then not necessary dynamic. Properties are computed by image processing routines whereas actions are described in natural language. Therefore works on describing action and behaviors in natural language are useful. As far as I am concerned, I have used works from the semantic community [8], [9], [7]. The following issue is how to deal with the gap between numerical properties (computed by image processing routines) and symbolic information (used by human operator). I have tried to answer this question in [4].

1.2 What can we learn from human perception of action?

Humans can usually recognize an action with little information (or a global information on the action). This is possible only because humans possess the context of the action.

I do believe this is also the case of a computer system for any type of action [1], [3]. For example, context can help with :

- selecting the right task to perform an efficient recognition,
- giving clues to guide the recognition task,
- making the link between numerical properties and symbolic information.

The main problem with context is its acquisition phase.

1.3 How do we represent action?

To represent the perception of actions, I use the notion of scenario. I recursively define a scenario as a combination of sub-scenarios. At level zero, a scenario is directly defined based on the properties of mobile objects in the scenario. Mobile objects can have two roles in a scenario : *source* or *target*. A source mobile object performs the action associated to the scenario and the target object is the reference of the action. A target object can be either static (e.g. an exit zone) or dynamic (e.g. a mobile object).

I also define two types of combination : *non temporal* or *temporal*. A scenario can correspond to a non temporal constraint on a set of sub-scenarios and it can also correspond to a temporal sequence of sub-scenarios. For example, the scenario *"the man is swaying"* represents a non temporal combination of two properties : the man trajectory is not straight and its shape is changing. The scenario *"the man drives up the car then he goes away"* represents a temporal combination of two sub-scenarios. I try to recognize a non temporal scenario through an abductive diagnosis [2] and to recognize a temporal scenario thanks to an automaton [5].

Several static hierarchies have already been defined for the recognition of car scenarios : elementary verbs at the base of the pyramid (e.g. *"to change of lines"*) and complex ones at the top (e.g. *"to overtake"*). I prefer recursive hierarchies, because we can adapt the scenario granularity to the application domain.

1.4 What is the role of "causal" or physics based descriptions in understanding action?

The recognition of complex static objects consists usually in a loop of two operations : (1) computing properties on the static object, (2) then classifying the object based on these properties and on a set of predefined object models. First, we start with generic properties and then we end up with specific ones. At each step, we hope to reduce the number of possible classes for the current object. Thus we can plan the computation of properties used to recognize static objects.

The recognition of actions differs in the fact that all properties cannot be computed at the beginning. We have to wait for other frames to compute new properties and to enhance the recognition. Instead of using classification or deduction, we have to use abductive diagnosis for action recognition. If we compute a property on an action, then it just gives a clue to generate a hypothesis to explain the property. Then we have to manipulate a set of hypotheses and to wait for additive properties to decide which one is true.

1.5 How should time be handled?

It depends on the application. If the properties on mobile objects are true/false at one time point and at one location, then we can use sophisticate time reasoning. For example in [6], they monitor robots in a plan. They use optical barriers as sensors, so they know at each time point robot locations. This is also the case of applications like the safety monitoring of planes in airports [10].

If you try to interpret human activities, you never know when a property starts or is really true. You have to deal with uncertainty and transitions between actions. At the present time, the recognition of actions is not robust enough to enable a complex manipulation of time. As far as I am concerned, I only use a very simple notion of time (e.g. How long has lasted the action?, Action 1 starts after action 2 finishes.).

1.6 What are good domains and example tasks to try to address?

- Traffic surveillance : accident prevention, convoy protection.
- Surveillance in metros, supermarkets, car-parks, gas-stations : prevention of vandalism.
- Sport : help for coaches with statistic information.
- Miscellaneous : intelligent room, video sequence retrieval.

In these domains, only simple tasks can be addressed at the present time. The best way is to start using image sequences of structured scenes, with few people (i.e. not a crowd), a good quality image (e.g. several sensors) and the application context.

2 My personal issues

2.1 How can be divided the global process of image sequence interpretation?

There are several hard tasks in image sequence interpretation : context acquisition, movement detection, tracking mobile objects, computing symbolic information from numerical properties, symbolic recognition of actions. A way to handle the relative problems consists in the cooperation of all these tasks. So one issue is also to divide the image sequence interpretation process into several sub-tasks and to define the connections between these sub-tasks. For example, How can work an action recognition module with a non robust tracking module? How the action recognition module can improve the tracking module?

2.2 What are the basic properties for action recognition?

In my opinion, the basic properties used to recognize an action are [4] :

- shape and shape evolution : outline, size, the color regions that compose a mobile object...
- trajectory and trajectory evolution : location, time, direction, speed...

These properties do not replace realistic interpretations, they just allow us to generate hypotheses on the action. They can be seen as clues to build the interpretation. The main point is not really what are the basic properties but rather how can we use them.

3 Expected goals in the M.P.A. workshop

First, I hope to have comments on my work and to be able to compare it to other works. I hope in particular that most of the participants will try to give their own view on the proposed topics.

Second, I also hope that constructive discussions can be done. Because the application domain is quite large, I think we may misunderstand our different view-points. When somebody is speaking, it could help if we have seen before some descriptions (a resume, few words?), images and video-tapes of the problem.

References

[1] F. Brémond and M. Thonnat. A context representation for surveillance systems. In *Proc. of Workshop on Conceptual Descriptions from Images at the European Conference on Computer Vision (ECCV)*, Cambridge, April 1996.

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

[3] F. Brémond and M. Thonnat. Issues in representing context illustrated by scene interpretation applications. In *proc. of the Int'l and Interdisciplinary Conf. on Modeling and Using Context (CONTEXT-97)*, Rio de Janeiro, February 1997.

[4] F. Brémond and M. Thonnat. Object tracking and scenario recognition for video-surveillance. In *the poster sessions of the 15th Int'l Joint Conference on Artificial Intelligence (IJCAI)*, Nagoya (Japan), August 1997.

[5] F. Brémond and M. Thonnat. Recognition of scenarios describing human activities. Submitted to the International Workshop on Dynamic Scene Recognition from Sensor Data, June 1997.

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

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

[8] 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.

[9] J. Schirra and E. Stopp. Antlima a listener model with mental images. *IJCAI, Chambéry, France*, 1:175–180, 1993.

[10] E. Tsang and R. Howarth. Scheduling in both space and time. In *11th Workshop on Expert Systems and their applications, Avignon*, pages 361–373, May 1991.