Issues of representing context illustrated by video-surveillance applications

François Brémond and Monique Thonnat
I.N.R.I.A., 2004 rte des Lucioles - BP 93
06902 Sophia-Antipolis Cedex, France
Tel : +33 93 65 76 59, Fax : +33 93 65 79 39
Email : francois.bremond@sophia.inria.fr

Keywords: Context modelling, knowledge representation, knowledge-based systems, scene interpretation, view-points.

Abstract

This paper tackles several issues of context representation in knowledge-based systems.

First, we propose a definition of context through the description of the different types of information manipulated by a process. Thanks to this definition we explain the role of the granularity level of processing and the role of the abstraction level of application in modelling context. Based on this definition two main issues related to context are tackled : how context representation can be built and organized and how context contents can be re-used for other applications. Then we propose several solutions to deal with these issues : using a multi view-point representation and describing context through symbolic information. We illustrate the proposed context model with the process of dynamic scene interpretation. After explaining the reasons why this process is particularly concerned with the use of contextual information, we describe the context representation and its implementation for this specific process. Finally we give an example illustrating the utilization of the context representation and we describe the software we have developed to ease the acquisition stage of context contents.

[0]International Journal of Human-Computer Studies Special Issue on Context

1 Introduction

The realization of a system which interprets a scene depicted by image sequences is an increasing research field. Even if the interpretation system is based on efficient algorithms, the results of the overall processing is not completely satisfactory due to the difficulty of making cooperate the different parts that compose this process. Several teams have developed approaches to overcome this problem. For example, we can quote H. Nagel [Nagel, 1988], J. Schirra [Schirra, 1990], A. Bobick [Bobick and Pinharez, 1995] and the esprit VIEWS project (Visual Inspection and Evaluation of Wide-area Scenes) [Corrall, 1992].

The solution we chose to improve the interpretation process is based on the utilization of contextual information, obviously important for such a process. However what is not clear is how to represent and use context. Although our first goal was the study of the context representation for the interpretation process, we have tackled general issues of context modelling in order to propose solutions which are independent of the application domain. Thus, this paper studies general issues related to context in knowledge-based systems : what is context, how context representation can be constructed and organized and how context contents can be re-used for other applications.

First we propose a generic context definition through the description of the different types of information manipulated by some process. This definition underlines the main characteristics of context representation and proposes methods to cope with them. Then we illustrate the context definition by describing the context for the interpretation process. We explain the interest of context utilization for this process and we describe its implementation.

2 Related Works

2.1 Multi-Disciplinary Issues

Many works have coped with the issue of context modelling. This issue is common to several disciplines and the importance of the notion of context is now widely acknowledged. Cognitive Scientists have long been discussing the meaning of context in an intellectual process. In Symbolic Reasoning several context formalizations have been proposed. For example in [McCarthy, 1993], the author defines context as a mathematical notion. He proposes rules to make context more generic and to provide an order relation for grading levels of context. In Computational Linguistic context is also a central notion helping in the understanding and the generating of texts written in natural language. For Knowledge-based systems contextual information eases the adaptation of systems to real-word conditions. For example in [Turner, 1995], the author uses context in Case-based reasoning. He has built a system that supervises the movement of an unmanned sub-aquatic vehicle and that uses context for making the vehicle react to unplanned events. When an event occurs like *"the vehicle arrives near the coasts"*, the system adapts the supervision of the vehicle through actions like *"to slow down"*. The main issue of this author is how to manage context automatically. In Computer Vision many systems use also context. For example in [Strat, 1993], the author enumerates several systems in object recognition that are able to reason using context. In particular the author is dealing with the issue of context representation in order to use context systematically [Strat and Fischler, 1990].

In all these disciplines the same issues are raising : how to define context, how to generate and use it automatically? Although no solution are really emerging the comparison of the works already accomplished in the different disciplines can help to answer fundamental issues in modelling and using context.

2.2 Local context

In the case of the scene interpretation there are different ways of using contextual information. A first approach consists in defining a context for each mobile object detected in the scene. This information is called local context and it is mainly used to ease the tracking process of mobile objects.

For example in [Intille and Bobick, 1995], the authors track football players using pieces of context, one for each player, called closed worlds. This local context contains the color distribution of the player and of its direct environment, like the marks on the ground. Thanks to this context the authors enhance the tracking of players knowing the surrounding static objects. In [Choi et al., 1997], the authors use also the color histogram of players and of their environment to track them and to cope with the dynamic occlusions due to other players. In [P. Remagnino and Kittler, 1993], the authors define a local context, called spatio-temporal context, which contains the life duration of a player and its location in order to supervise the mobile CCD camera used to observe the scene.

Local context can be used also to improve the recognition of actions performed by mobile objects. For example in [Nagel, 1988], the author defines a context as an information combining generic descriptions of spatio-temporal structures and of the intention of vehicles moving in a scene corresponding to a road. The author plans to use this notion of context to recognize actions like *"to park a car"*. However he does not describe how he is going to represent and use contextual information.

Therefore local context contains usually numerical information describing the environment surrounding one mobile object. This information is mainly used to improve the tracking process of the mobile objects.

2.3 Mobile sensor and multi-sensor systems

Systems equipped with several sensors or with a mobile sensor often use context in order to control sensors and their resources.

For example in [Ghallab et al., 1992], the authors are mainly interested in modelling context relative to sensors and to perceived data. They have built a mobile robot equipped with several sensors and able to understand its environment. The word *"sensor"* represents here a physical sensor (e.g. a CCD camera) as well as an image processing program (e.g. a color region extractor). The description of a sensor is composed of three models : a structural model, a state model and a perceptual model. The structural model defines the type of data generated by the sensor and can be recursively defined thanks to the structural models of lower level sensors. The state model contains a set of variables characterizing the sensor, like the location or the focal length of a camera. The perceptual model describes the relationships between the data generated by the sensor and their meaning in the scene. For example this model gives the probability of the detection of primitives computed by image processing programs, like a 2D line segment. All this contextual information is used to control sensors, to manage their resources, to schedule image processing programs and to take into account the geometry and the reliability of sensors.

In [Clement et al., 1993], the authors have also developed a multi-sensor system able to reason and use context. This system has to analyze a scene observed by a satellite. The authors adapt the recognition of the scene objects (e.g. a bridge) depending on sensors. Thus this contextual information allows the system to adapt the perception of data to the current scene.

2.4 Map of the scene

In scene interpretation the context the most widely used concerns information on the spatial structure of the scene, usually called the scene map. We present below several systems using such a map :

- First in [Neumann, 1984], [Mohnhaupt and Neumann, 1990], the authors have developed the system Naos which from an image sequence describes a scene of a road to a listener unable to see the scene. Naos uses a geometric description of the scene containing photometric properties (e.g. color, illumination). This description corresponds to a 2D map of the scene, composed of a cell set structured like a grid. The authors say that this representation of space is analogical because these cells correspond explicitly to the inherent structure of the scene. Thanks to this representation the system Naos reasons directly on the 2D map instead of reasoning on the image sequence. The spatio-temporal relations of mobile objects are computed at the level of cells. For example the distance between a vehicle A and a vehicle B is computed by sending a message from the cell containing A to the neighboring cells, in order to question them on the presence of B. Therefore Naos uses context to improve the computation of spatio-temporal relations and more especially as a reasoning support for the interpretation process.

- In [Tsuji and Li, 1993], the authors have built a map representing the panoramic view of an outdoor environment corresponding to a road lined with buildings. To spare memory they only keep on the map the main buildings, called landmarks. The authors use this contextual information for a robot moving in a scene to be able to find its way around.

- In [Sellam and Boulmakoul, 1994], [Cerf and Pintado, 1997], the authors have defined sophisticated models of a scene corresponding to a connected set of crossroads, in order to supervise the pedestrian and vehicular traffic. These models contain topologic information, like nearby junctions, links, entry and exit sections. They also contain information on the traffic signals, such as the location and the state of traffic lights, traffic phases and pedestrian signals. All this contextual information allows the interpretation system to track mobile objects and to establish the traffic flow.

- In [Bobick and Davis, 1996], the authors have developed a system that supervises television cameras in a studio. This system is characterized by an approximate model of the world, continuously updated during the processing. This model contains the spatial structure of the scene with the location and the viewing angle of the cameras. It also contains rules (post and pre-conditions) helping in supervising image processing programs. A rule example is : IF *the mobile object is located in the image center* (pre-condition), THEN *extract the moving region in the center* (action) AND *the mobile object and the extracted moving region should have the same surface area* (post-condition). Thus the authors use an approximate model of the world (a coarse 3D map) associated with rules allowing the system to understand the scene.

- The goal of the European Esprit VIEWS project is the real-time surveillance of outdoor scenes, based on the analysis of image sequences [Duong et al., 1990a], [Duong et al., 1990b]. This team extends the analogical representation of the scene proposed by B. Neumann, by defining an arborescent hierarchy of cells [Howarth and Buxton, 1992]. This new structure allows the VIEWS project to define zones at several abstraction levels, including symbolic information relative to these zones. For example, the *"road"* zone contains a child zone *"traffic light"*, indicating that the stop of a vehicle in this zone could be due to the presence of a traffic light. This information is used as a constraint to check the consistency of the system at all the processing levels. In the VIEWS project contextual information is then widely used from image processing level to the abstract level of behavior recognition.

All this contextual information is mainly used in order to adapt the perception of data to their location in the scene environment. Context is usually represented as a map of the scene and is widely used from lower levels (e.g. computation of spatial relations) to more abstract levels (e.g. behavior analysis).

2.5 Building context automatically

In computer vision there are many systems enabling to build automatically context for a given static scene. The main goal of these systems is to determine the spatial structures of the scene (e.g. the arrangement of buildings). In [Milhaud and Médioni, 1994], [Nevatia and Médioni, 1996], the authors present several systems dedicated to the reconstruction of static scenes. In particular they propose to build the model of a site like a plan from aerial images. They extract line segments from an image and group them to construct the model of the buildings.

These systems show that under some conditions it is possible to generate the spatial and structural information of a static scene. However in dynamic scene interpretation only few systems have a pre-processing stage dedicated to the generation of the scene context. In the best case, this generation consists in the construction of an approximate map of the scene.

3 Context modelling

In general speaking the scientific community acknowledges the importance of context. However the effective utilization of context remains elusive, mainly due to the difficulty to formalize this notion. Thus defining what we mean by context is a crucial issue.

3.1 Definition

The definition of the process context depends on the process nature. For H. Nagel [Nagel, 1988] the context of an action analysis process is a complex structure, comprising generic descriptions for spatial structures, temporal changes associated with these structures and the intention aspect of the action. For T. Strat [Strat, 1993], the context of an image understanding process is, in the broadest sense, any and all information that may influence the way a scene is perceived. More generally, a process uses three types of information : knowledge, contextual information and factual information. **Knowledge** is always valid. It is directly connected with the process goals and often belongs to a well defined model. If the process misses one part of knowledge, it is no longer able to compute results. **Contextual information** depends on applications, but it remains constant during processing. It is an accessory information, but it may become essential to handle particular situations. Context enables to improve the processing and corresponds to the additional information needed by the process to work efficiently. **Factual information** depends on the processing states. Its life time is often short. It corresponds to input data and computed data. Thus we propose to define the contextual information of a process as the information verifying two conditions:

$$\begin{array}{l} \text{- its value remains constant during processing,} \\ \text{- its value changes when the process is used for another application.} \end{array} \quad (1)$$

This definition of context has two main consequences. First we have to choose from which granularity level of processing we consider an information as factual rather than contextual. Second we have to choose from which abstraction level of application we consider an information as contextual rather than belonging to knowledge. The difficulties of formalizing the notion of context come mainly from the dependency of context to the application domain. These difficulties come also from the fuzzy border separating context from the other types of information. For these reasons the literature contains very few formal definitions of context. However this definition is necessary as soon as we plan to rationalize the utilization of context.

3.2 Interpretation process

In this paper we illustrate the proposed context definition through the example of the dynamic scene interpretation process. More particularly the class of applications we are interested in, is the automatic interpretation of indoor and outdoor partially structured scenes with a fixed monocular color camera. Given image sequences describing a scene, an interpretation system has to identify the behaviors of mobile objects. In our case, mobile objects can be either humans or vehicles. They constitute the moving scene objects, as opposed to static objects that belong to the static environment. By interpretation, we mean the overall interpretation process from image processing to behavior analysis. This process can be divided in three main tasks, which can overlap each other: **moving region detection**, **mobile object tracking** and **behavior analysis**. From images, the moving region detection task detects the movement of mobile objects thanks to image processing programs. Then the mobile object tracking task associates the detected moving regions to form and track mobile objects. Finally, the behavior analysis task interprets event occurrences concerning mobile objects and analyzes their behaviors.

3.3 Several granularity levels of processing

The definition 1 indicates that for defining the context of the interpretation process we have to choose the granularity level of this process. As said before, the interpretation process is composed of three main tasks. In addition a basic task, **spatial reasoning**, is used by two of the main tasks : the object tracking and behavior analysis tasks. Thus we consider that four specific tasks compose the interpretation process, which are the three main interpretation tasks and the spatial reasoning task. So the interpretation process can be considered at two different levels of processing : at the level of the global process or at the level of the four tasks. We choose to define context according to the tasks, because each task belongs to a well defined knowledge domain. Therefore we define the context of the global interpretation process as following (*task-i* denotes any task of the interpretation process) :

$$\text{context[global process]} = \bigcup_i \text{context[task-i]} \tag{2}$$

This definition provides us with a rule that can decide whether a piece of information verifies the first condition of definition 1 : if it remains constant during the processing of at least one task that uses it, then the piece of information is not factual.

3.4 Several abstraction levels of application

To define precisely the context of a process, definition 1 indicates that we have to choose the abstraction level of application in order to decide on which information domains the process depends. For the interpretation process, we can consider an application at the general level of video-surveillance, at the more specific level of metro station surveillance or at the even more specific level of the surveillance of a specific station observed by a specific camera.

In this paper the considered abstraction level of application corresponds to the surveillance of a specific metro station. So once this level is determined we can take inventory of the sources (as suggested in [Strat, 1993]), where contextual information may come from :

- **Scene Environment Information (SEI)** - It embraces spatial structures (e.g. tessellations, calibration plans), static objects (e.g. pillars, escalators), optical characteristics (e.g. reflections on the ground, occlusions) and behavior characteristics (e.g. exit and cluttered zones, roads).
- **Image Acquisition Information (IAI)** - It includes camera characteristics (e.g. camera model, focal length, color, filter), image characteristics

(e.g. image size and type, date and time of acquisition) and acquisition characteristics (e.g. camera orientation and camera position).

- **Derived Temporal Information (DTI)** - It is obtained as a result of earlier executions of interpretation tasks. This information can be seen as the accumulated information about the past and the prediction about the future (e.g. already detected and identified mobile objects).
- **User Request Information (URI)** - In an interactive way, a human operator can provide context to the overall system during processing. A typical example for surveillance system is an operator request for detecting specific persons.

The interpretation process depends on all these source domains of contextual information. This property provides us with a rule that can decide whether a piece of information verifies the second condition of definition 1 : if it belongs to one of these source domains, then the piece of information is not considered as knowledge.

3.5 Multi view-point context

When we try to represent the context of a process we often have to deal with two general issues : how context can be built and organized and how its contents can be re-used for other applications. In interpretation applications these two issues have to be tackled.

3.6 Space representation

A first problem is due to the large size of context domains. The interpretation process is composed of several tasks, each task using widely contextual information. This information is then spread all over the interpretation system, making difficult a centralized representation of context.

A second problem comes from the choice of definition 2 for the interpretation process context. Because of definition 2 we consider that all the context of a task, even provided during the processing of another task, belongs to the global process context. This case raises with the context domain of Derived Temporal Information (DTI). For example the degree of interest of a mobile object is computed by the behavior analysis task but it is also useful for the moving region detection task to select on the next image the regions of interest. So the detection task needs this information to improve its computation. Therefore a mechanism is needed to share the common information between interpretation tasks. The solution we choose is to use a centralized representation of context, with a uniform formalism for all interpretation tasks. We propose to use the representation of space as a support for the context representation for two reasons:

- For the interpretation process the representation of space is a basic structure common to two main tasks.
- The context of the interpretation process comes mainly from the context domain of Scene Environment Information (SEI).

Thus the representation of space is selected to gather all context domains in one place. As presented in the state of the art, this solution has been first proposed by [Mohnhaupt and Neumann, 1990]. These authors have defined a map of the scene to improve the computation of spatio-temporal relations relative to mobile objects. By this way the representation of space can be seen as a reasoning support for the interpretation process and eases the organization of the context representation.

3.7 Multi view-point representation

By definition context depends on applications and so its contents have to be acquired before its utilization. The issue of acquisition is the main issue in context utilization. In our interpretation system contextual information is provided by human operators. Since four task context domains are defined for the interpretation process, the context acquisition requires several operators, at least one for each interpretation task. So another problem is the mix of the acquisition of all task context domains. This problem becomes crucial when large systems are developed. The solution we propose, is to use a multi view-point representation. In this representation a view-point is a way of seeing context for a specific interpretation task by filtering the corresponding element of context. For interpretation systems four view-points are needed, one for each interpretation task. Thus view-points divide context representation and help operators with its construction.

Therefore the context representation allows us to divide context during its construction phase through view-points and to centralize and share context during the processing phase through the representation of space.

3.8 Re-used of context

Another approach to ease the context acquisition phase consists in re-using context that has been already acquired. The issue of re-using old contextual information is all the more relevant that a high abstraction level of application is considered. If a high abstraction level is chosen, then the process depends on several source domains of contextual information, making the context contents numerous.

For the interpretation process context depends on four source domains of contextual information. Thus a big amount of contextual information has to be acquired, especially because of the context source domain of Scene Environment Information (SEI). A solution is to automate the acquisition phase, but this issue is particularly difficult when information on human activities has to be acquired. The solution we choose is to re-use the contextual information generated for other applications. As suggested in [McCarthy, 1993], a main issue in context formalization is to abstract information. In our representation this idea is realized by the utilization of symbolic information. Instead of representing context with a complete description, only symbolic information is used. The complete description is predefined in libraries belonging to the interpretation system. For example we just use symbols for the description of the static objects of the scene environment and for the description of behaviors. Thus one part of the interpretation context is generic and can be used for different applications (e.g. for the surveillance of different metro stations).

The proposed solutions to ease the acquisition phase come mainly from software engineering. They show that the utilization of context is possible, but it is not effortless.

4 Context representation

4.1 Decomposition of space

As previously explained, we propose to represent context through the representation of space. For an interpretation system the space corresponds to the projection of the 3D scene on the 2D image plan. We decide to represent the space by a decomposition of the 2D image plan into a partition of zones delimited by polygons. We choose polygons because they are structures simple enough to be easily implemented, and flexible enough to match human knowledge. The polygonal zones have to be drawn off line by a human operator. This drawing operation is a necessary stage before using the interpretation system. Each polygonal zone is linked to four context elements. We call **context element** pieces of contextual information associated to a zone and to an interpretation task. A zone delimits the location where its corresponding context elements can be applied. Figure 1 shows the polygonal zone *"Z1"* with its four corresponding context elements ranged from *"spatial-reasoning-context-element-1.1"* to *"behavior-analysis-context-element-1.4"*. For example the tracking task can use the *"mobile-object-tracking-context-element-1.3"* to improve the tracking of objects moving in the zone *"Z1"* during all the processing of the image sequence. Therefore the **decomposition of space** is the support of context representation and centralizes all its contents. At the present time we just have defined one level of zones. In our futur works we plan to construct a hierarchy of zones to structure contextual information as it was suggested in [Howarth and Buxton, 1992].

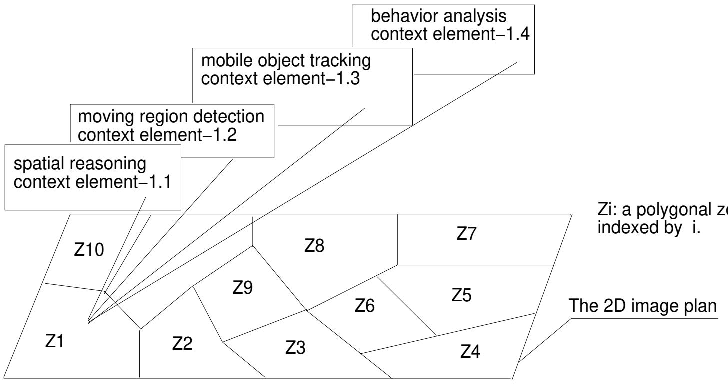

Figure 1: A context element for each interpretation task is linked to a polygonal zone of the decomposition of space.

A context element is defined as a sub-set of pieces of contextual information. It is represented thanks to the *"frames"* formalism : its slots correspond to the context properties. We have defined four classes of context elements, one for each interpretation task. The structure of a context element depends on the task and its values depend on the associate zone. This context representation verifies the constraints specified in the previous section. Context is centralized at one location with a common formalism. It is also divided through four view-points defined by the context element classes. Thus this representation eases the acquisition phase of context and its utilization.

The context of the interpretation process is then represented through a set of context elements. The remaining part of this section gives some examples of context element type and contents for each interpretation task.

4.2 Context in spatial reasoning

Spatial reasoning is a subpart of both the mobile object tracking and behavior analysis tasks. Its goal consists in computing spatial relations between all mobile objects. Then the spatial reasoning task needs context in order to enhance the computation of these relations.

A first problem related to this computation raises when the interpretation system analyzes numerous mobile objects. For example, the computation of the neighbors of a given object implies an access to all object positions. To reduce complexity, a common method suggested in [Mohnhaupt and Neumann, 1990] is to link each mobile object to the zone it occupies. After indexing all objects to the set of polygonal zones belonging to the decomposition of space, spatial relations are computed through the zones instead of the object positions. For example, the neighbors of a given mobile object can be computed just by accessing to the zones adjacent to the one occupied by the object. As spatial reasoning is a basic process for the interpretation system, the utilization of the decomposition of space improves significantly the overall system performances.

A second problem is due to the nature of image processing data which are computed in the 2D image plan. To get accurate relations, several features of mobile object need 3D computation. For example two mobile objects, which are closed to each other on the 2D image plan, may be far away in reality as soon as their depth in the scene differs. The spatial reasoning task needs therefore to transform properties from the 2D image plan to properties belonging to the 3D scene. A solution consists in using the context domain of Scene Environment Information (SEI) and for example to use a calibration matrix. This matrix allows the interpretation system to compute the 3D scene coordinates of a mobile object depending on its 2D image plan coordinates. Usually in this case, an accuracy coefficient is used to establish the accuracy of the 3D scene coordinate computation.

For a given zone we define the spatial reasoning context element thanks to three slots : the calibration matrix, the accuracy coefficient of the 3D coordinate computation, and the link to the corresponding polygonal zone. Thus, the spatial reasoning task can take advantage of contextual information in two different ways, by improving the computation time of basic spatial relations and by performing accurate 3D reasoning.

4.3 Context in moving region detection

The detection of mobile regions and more generally image processing programs are complex tasks for several reasons. Often in image processing, a single algorithm cannot solve alone a given problem : several steps are involved to compute the final results and each step may be realized in different ways. Thus image processing programs need program supervision [Thonnat et al., 1994] (i.e. need to be selected, scheduled and linked to each other). Moreover, image processing parameters need tuning during execution before obtaining satisfactory results. Finally, most of interpretation systems must work in real-time, making crucial the focus of system attention to specific image regions of interest [Buxton and Gong, 1995].

To supervise image processing programs, systems use traditionally the context domain of Image Acquisition Information (IAI). It allows the interpretation system to obtain more accurate results by tuning the processing according to application particularities and to new scene conditions. It can be used directly as parameter values for algorithms, or for both planning and control tasks in order to use programs in situations for which their designers intended to use them [Moisan et al., 1995], [Strat, 1993]. The context domain of Scene Environment Information (SEI) may also be used to satisfy these goals, by adapting image processing programs to specific areas of the scene. When image processing programs are part of an interactif system, the context domain of User Request Information (URI) is often used to take advantage of the user knowledge. For example this context allows the system to select the level of details of the processing.

As in our interpretation system we use image processing programs that do not handle scene characteristics, we define the same mobile region detection context element for all the polygonal zones of the decomposition. This context element contains global slots like the size of one image and the frequency of the image sequence. This use of context helps to supervise the processing and to increase the results quality of the detection task.

4.4 Context in mobile object tracking

The mobile object tracking task has to form and track scene objects. It gathers the moving regions detected by image processing programs in order to form new objects and to match these new objects with the ones already tracked. The main problem of the tracking task comes from the bad quality of detected moving regions due to optical irregularities, such as reflections on the ground, shadows, cluttered zones, blinking lights, occlusions, and patterned backgrounds. To gather the moving regions, the tracking task has to generate assumptions that influence the remaining part of the interpretation process. Even with precise models describing the geometric structure of mobile objects, the matching process is uncertain [Grandjean, 1991], [Koller et al., 1993]. This situation becomes worse when no precise mobile object model is available, like in the case of badly detected human beings. Therefore the computed mobile objects are often false or erroneous, and previously tracked objects may be lost during the tracking process. A usual way to handle this problem is to compute the uncertainty of assumptions [Ghallab et al., 1992] or to compute the probabilities of assumption validity which can be used in Bayesian networks [Nicholson, 1992], [Buxton and Gong, 1995].

To handle this issue, the tracking task can take advantage of the context. First thanks to the context domain of the Scene Environment Information (SEI), the interpretation system can know the existence of an optical irregularity in a particular zone. It can then apply a specific processing on the moving regions of the zone before matching them with the mobile objects already tracked. Thanks to context, the interpretation system can also know specific properties on the given zone concerning the management of objet tracks. For example, a zone can be labeled as an exit zone allowing the system to expect the ending of some mobile object tracks. From all this information, the interpretation system can also deduce the uncertainty of the tracking process. For example, a mobile object belonging to a zone that holds numerous optical irregularities is likely to be uncertain. The computation of uncertainty is a solution to provide *a priori* probabilities for Bayesian networks. Second, the mobile object tracking task can use the context domain of the Derived Temporal Information (DTI) to track mobile objects. For example, the interpretation system can predict the new location of an already tracked object and focuses its attention to the expected location. So using context is essential for mobile object tracking and for establishing the confidence in the whole interpretation system.

Thus for a given polygonal zone we define the mobile object tracking context element thanks to four slots : the optical irregularity descriptions, specific properties on zones relative to object tracks, the uncertainty coefficient of the tracking process and the link to the corresponding polygonal zone. This zone is supposed to delimit the optical irregularity influence. This definition implies that a library of optical irregularities containing complete descriptions is predefined in the interpretation system. This utilization of a predefined library allows us to represent context with symbolic information instead of complete descriptions. Therefore the context element associated to the mobile object tracking task can be generated more easily.

4.5 Context in behavior analysis

The role of the behavior analysis task is to link the numerical properties of mobile objects to behavior models expressed in natural language. This role consists first in abstracting the object properties and then in selecting a model of behavior explaining the properties. What makes the behavior analysis task so difficult is the big gap between behavior models and object properties. The properties have numerical values which are often instantaneous. On the opposite behaviors are usually expressed in natural language, are dependent on applications and are defined on long image sequences. The issue of linking numerical properties to behavior models is a common one for all systems integrating natural language and visual data [Srihari, 1994]. There are two ways of handling this problem. A first solution is to accurately describe object behaviors in natural language. A common method is to use spatio-temporal propositions to precisely describe behaviors, but both quantitative and qualitative problems have to be tackled [Olivier et al., 1994]. A second method is to abstract object properties. For example in [Nagel, 1988], the author classes motion verbs into a taxonomy of increasingly complex actions ranging from object properties to abstract activities like parking a car.

A way to improve these methods is to use contextual information as a bridge between the properties of mobile objects and their behaviors. The context of the analysis task can be seen as a set of links. As underlined by Howarth [Howarth, 1995], most areas are named either by their location (e.g. a kitchen) or by the actions that occur in them (e.g. cooking), providing to the analysis task important clues on expected behaviors. This information is contained in the behavior characteristics of the context domain of Source Environment Information (SEI).

For a given polygonal zone we define the behavior analysis context element with four slots : the static objects of the environment belonging to the area (e.g. a seat), expected properties on mobile objects (e.g. the speed limit is low), expected behavior models (e.g. to seat down) and the link to the corresponding polygonal zone. As remarked before, this context element description implies that libraries of static objects, of object properties and of behavior models are predefined in the interpretation system. Thanks to this context element the interpretation system is then able to enhance the behavior analysis task.

Therefore, the four interpretation tasks can take advantage of context in several ways. For complex applications, this context utilization is even a crucial issue.

5 Using the context representation

To illustrate the benefits of the context representation we describe one of its typical utilization during the execution of the interpretation process. In this example, we use an image sequence of a metro station in Charleroi which has been taken by SRWT in the framework of the Esprit PASSWORDS project (Parallel and real time Advanced Surveillance System With Operator assistance for Revealing Dangerous Situations) [Chleq and Thonnat, 1996]. This image sequence depicts a man moving behind a seat. The goal of the interpretation system is to track the man even if he is occluded by the seat.

In the image center of figure 2, the man is partially occluded by the back of the seat. Three moving regions are detected, but we are just interested in the two regions in the middle of the image. First for these two moving regions, the interpretation system retrieves the polygonal zones that contain the regions. As the zones are linked to mobile object tracking context elements, the interpretation system is aware of the potential presence of an occlusion and then it is able to apply a specific processing on every moving region belonging to these zones in order to handle the optical irregularity. Knowing the borders of the seat back thanks to the context elements, the interpretation system tries to gather the top moving region together with the moving region just below. So it merges the two regions assuming that they represent a unique mobile object partially occluded. Then using again the context elements, the interpretation system establishes an uncertainty coefficient for the tracking of the mobile object, according to the object size and to the characteristics

Figure 2: Frame 1 is a raw image taken from the image sequence. Frame 2 is the result of the moving region detection task. It shows two moving regions partially occluded by the back of a seat.

of the optical irregularity. Therefore thanks to context, the tracking task is able to track correctly the mobile object even partially occluded, allowing the interpretation system to continue the behavior analysis of the man.

6 A context acquisition software

Despite the utilization of symbolic information context acquisition is still a tiresome phase. For this reason we have developed a graphical user interface, allowing a human operator to acquire interactively contextual information and to construct the context representation of a scene. It is implemented in C language and uses Motif toolkit [Brémond and Thonnat, 1996]. Figure 3 shows a polygonal zone drawn by a human operator and one of its corresponding context elements. As soon as the polygon is drawn, the software asks the operator for the values of all the slots of the corresponding context element. Each complex slot is represented by a symbolic value linking the zone to predefined libraries. For example the value *"seat"* of the slot *"static objects"* links the current zone to a complete seat description predefined in the interpretation system. Thus, the interface creates a context representation for a specific scene of a given application. As all complex descriptions are symbolic, the context contents can be partly re-used for another scene of the same application domain. The context representation can also be used by any interpretation

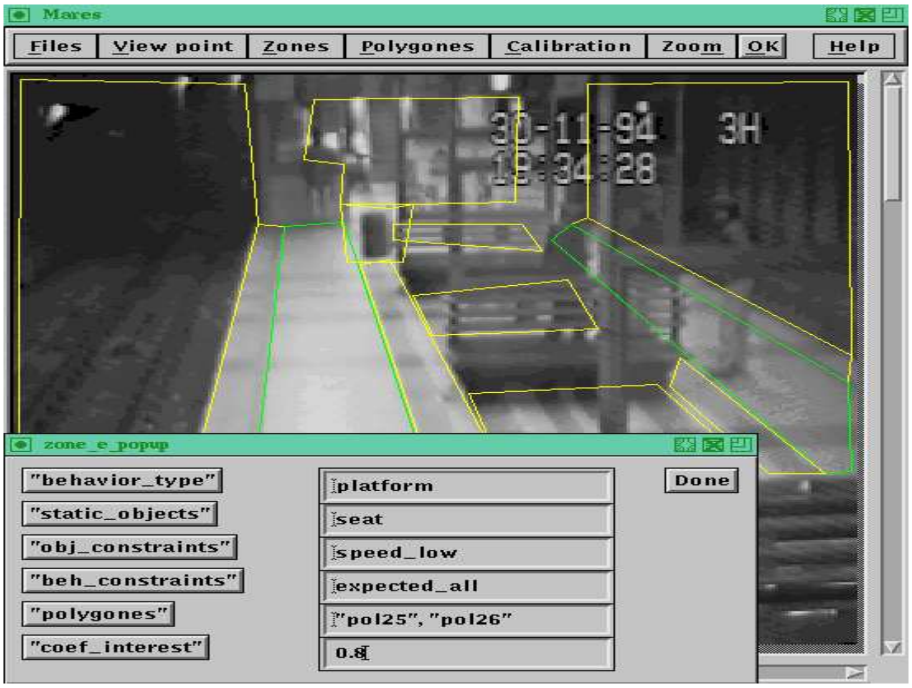

Figure 3: A human operator is using the interface to fill the values of the behavior analysis context element for the polygonal zones in the ground (polygonal zones are drawn in white).

system that possesses the same symbolic description of context.

7 Conclusion

This work proposes a general definition of context for knowledge-based systems through the description of different types of information. This definition depends on the granularity level of the processing and on the abstraction level of the application. The context definition allows us to define the context of a complex process such as the process of dynamic scene interpretation. It allows us also to underline the role of the sources where contextual information may come from. Deduced from the context definition, two rules enable us to delimit the scope of context.

Then, we have tackled two general issues related to the construction of context representation and to the re-used of context contents for other applications. To handle these issues we propose in one hand to divide context thanks to a multi view-point representation based on the frame formalism and in the other hand to centralize the whole context at one place thanks to the decomposition of space. So our representation allows us to see context under different view-points and at the same time to share common contextual information between all interpretation tasks. We propose also a way to abstract contextual information in order to obtain a partially generic context representation and to enable the re-use of its contents. Thus context representation becomes easier to construct, to acquire and to use.

However we are also interested in applications using mobile sensors that modify the perception of context during the processing. For example we plan to interpret scenes observed by a mobile camera located on a plane. Our future works will consist then to adapt the context representation in order to generate and update automatically contextual information during the interpretation processing. For example this automation can be realized thanks to statistical methods or learning technics.

References

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

[Bobick and Pinharez, 1995] Bobick, A. and Pinharez, C. (1995). Using approximate models as source of contextual information for vision processing. In *proc. of the IEEE workshop on Context-Based Reasoning (ICCV'95).*

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

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

[Cerf and Pintado, 1997] Cerf, V. L. and Pintado, M. (1997). An adaptive model of camera-driven urban intersections observation. In *Proc. of the*

International Workshop on Dynamic Scene Recognition from Sensor Data, Onera-Cert, Toulouse.

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

[Choi et al., 1997] Choi, S., Seo, Y., Kim, H., and Hong, K. (1997). Where are the ball and players? Soccer game analysis with color-based tracking and image mosaik. In *ICIAP'97*. to appear.

[Clement et al., 1993] Clement, V., Giraudon, G., Houzelle, S., and Sandakly, F. (1993). Interpretation of remotely sensed images in a context of multisensor fusion using a multispecialist architecture. *IEEE Transaction on Geoscience and Remote Sensing*, 31(4):779–791.

[Corrall, 1992] Corrall, D. (1992). Deliverable 3: Visual monitoring and surveillance of wide-area outdoor scenes. Technical report, Esprit Project 2152: VIEWS.

[Duong et al., 1990a] Duong, V., Buxton, H., Howarth, R., Toal, P., Gong, S., King, S., Thoméré, J., and Hyde, J. (1990a). D203: Spatio-temporal reasoning. Technical report, Esprit Project 2152: VIEWS.

[Duong et al., 1990b] Duong, V., Howard, R., Hill, G., Toal, P., King, S., Gong, S., Thoméré, J., and Hyde, J. (1990b). D201: The representation of event, behaviour and scene. Technical report, Esprit Project 2152: VIEWS.

[Ghallab et al., 1992] Ghallab, M., Grandjean, P., Lacroix, S., and Thibault, J. (1992). Représentations et raisonnement pour une machine de perception multi-sensorielle. In *Proc. of the PRC-GDR IA, Marseille*, pages 121–167.

[Grandjean, 1991] Grandjean, P. (1991). *Perception multisensorielle et interprétation de scènes*. PhD thesis, LAAS - Université Paul Sabatier de Toulouse.

[Howarth, 1995] Howarth, R. (1995). Interpreting a dynamic and uncertain world: high-level vision. *Artificial Intelligence*, 9(1):37–63.

[Howarth and Buxton, 1992] Howarth, R. and Buxton, H. (1992). Analogical representation of space and time. In *Alveys conference*, volume 10, pages 467–478.

[Intille and Bobick, 1995] Intille, S. and Bobick, A. (1995). Closed-world tracking. In *Proc. of the 5th Int'l Conference on Computer Vision (ICCV)*, Cambridge, MA.

[Koller et al., 1993] Koller, D., Daniilidis, K., and Nagel, H. (1993). Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10(3):257–281.

[McCarthy, 1993] McCarthy, J. (1993). Notes on formalizing context. In *Proc. of the IJCAI, Chambery (France)*, pages 555–560.

[Milhaud and Médioni, 1994] Milhaud, N. and Médioni, G. (1994). Learning, recognition and navigation from a sequence of infrared images. In *Proc. of the Int'l Conf. on Pattern Recognition (ICPR)*, volume 1, pages 822–825, Jerusalem.

[Mohnhaupt and Neumann, 1990] Mohnhaupt, M. and Neumann, B. (1990). Understanding object motion: Recognition, learning and spatiotemporal reasoning. Research Report FBI-HH-B-145/90, University of Hamburg.

[Moisan et al., 1995] Moisan, S., Shekhar, C., and Thonnat, M. (1995). Real-time perception program supervision for vehicle driving assistance. In *proc. of the Int'l Conf. on Recent Advances in Mechatronics, ICRAM'95*, volume 1, Istanbul, Turkey.

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

[Neumann, 1984] Neumann, B. (1984). Natural language description of time-varying scenes. Technical report, FBI-HH-B-105/84 Fachbereich Informatik der Universität Hamburg, FRG.

[Nevatia and Médioni, 1996] Nevatia, R. and Médioni, G. (1996). Computer vision research at the university of southern california. *Int'l Journal of Computer Vision*.

[Nicholson, 1992] Nicholson, A. (1992). *Monitoring Discrete Environments Using Dynamic Belief Networks*. PhD thesis, University of Oxford.

[Olivier et al., 1994] Olivier, P., Maeda, T., and Tsujii, J. (1994). Automatic depiction of spatial descriptions. In *Proc. of the AAAI Seattle, Washington*, pages 1405–1410.

[P. Remagnino and Kittler, 1993] P. Remagnino, J. Matas, J. I. and Kittler, J. (1993). A scene interpretation module for an active vision system. In *SPIE*, volume 2056, pages 98–107.

[Schirra, 1990] Schirra, J. R. J. (1990). A contribution to reference semantics of spatial prepositions: The visualization problem and its solution in vitra. Research Report 75, VITRA-report.

[Sellam and Boulmakoul, 1994] Sellam, S. and Boulmakoul, A. (1994). Intelligent intersection: Artificial intelligence and computer vision techniques for automatic incident detection. *Artif. Intell. Applic. to Traffic Engng*, pages 189–200.

[Srihari, 1994] Srihari, R. (1994). Computational models for integrating linguistic and visual information: A survey. *AIR*, 8(5-6):349–369.

[Strat, 1993] Strat, T. (1993). Employing contextual information in computer vision. In *DARPA93*, pages 217–229.

[Strat and Fischler, 1990] Strat, T. and Fischler, M. (1990). A context-based recognition system for natural scenes and complex domains. In *DARPA90*, pages 456–472.

[Thonnat et al., 1994] Thonnat, M., Clement, V., and van den Elst, J. (1994). Supervision of perception tasks for autonomous systems: the OCAPI approach. *Journal of Information Science and Technology*, 3(2):140–163.

[Tsuji and Li, 1993] Tsuji, S. and Li, S. (1993). Making cognitive map of outdoor environment. In *Proc. of the IJCAI*, pages 1632–1638.

[Turner, 1995] Turner, R. (1995). Context-sensitive, adaptive reasoning for intelligent AUV control: Orca project update. In *proc. of the 9th Int'l Symposium on Unmanned Untethered Submersible Technology (AUV'95)*, New Hampshire.