Integration of vision and decision-making in an autonomous airborne vehicle for traffic surveillance

Silvia Coradeschi*, Lars Karlsson*, Klas Nordberg+

*Department of Computer and Information Science
+Department of Electrical Engineering
Linköping University, Sweden
E-Mail: silco@ida.liu.se, larka@ida.liu.se, klas@isy.liu.se

Abstract. In this paper we present a system which integrates computer vision and decision-making in an autonomous airborne vehicle that performs traffic surveillance tasks. The main factors that make the integration of vision and decision-making a challenging problem are: the qualitatively different kind of information at the decision-making and vision levels, the need for integration of dynamically acquired information with a priori knowledge, e.g. GIS information, and the need of close feedback and guidance of the vision module by the decision-making module. Given the complex interaction between the vision module and the decision-making module we propose the adoption of an intermediate structure, called Scene Information Manager, and describe its structure and functionalities.

1 Introduction

This paper reports the ongoing work on the development of an architecture for Unmanned Airborne Vehicles (UAVs) within the WITAS project at Linköping University. One of the main efforts within the project has been to achieve an efficient integration between a vision module, dedicated to tasks such as object recognition, velocity estimation, camera control, and an autonomous decision-making module which is responsible for deliberative and reactive behaviors of the system. A critical issue in such a system is to handle the fact that the vision module represents an object in terms of coordinates in some reference frame, whereas the decision-making module represents the same object in relational and qualitative terms. For example, the vision module can represent a car as a point in the image together with some parameters that describe its shape. The decision-making module, on the other hand, will represent the same car in terms of its relation to some road, or to other cars, and describe its shape in terms of symbolic attributes rather than as estimated parameters.

A second issue to be handled in the project is the integration of a priori information, here referred to as static information, and dynamically acquired information, e.g., produced by the vision module. An example of this integration

is how to combine information about the shape and topology of a road network, stored in a conventional GIS (Geographic Information System), and descriptions about cars (position, shape, etc) produced by the vision system, assuming that these cars are moving along the roads.

Section 2 presents a more thorough discussion on these and other issues. The general conclusion is that we need an intermediate structure, the Scene Information Manager (SIM), located between the vision and the decision-making module. The SIM solves both the problem of translating object references, e.g, from image coordinates to symbolic road labels, and vice versa, as well as manages the linkage of dynamic to static information and high-level prediction. The structure and functionalities of the SIM are described in more detail in section 3.

The resulting architecture has been implemented and tested on a number of scenarios. Section 4 briefly presents some of them and describes how the SIM is used to solve the above integration issues, thereby allowing the system to maintain a suitable distinction in abstraction level between the task driven vision module, mainly devoted to low-level vision processing, and the decision-making module which operates on symbolic information.

1.1 The WITAS project

The WITAS project, initiated in January 1997, is devoted to research on information technology for autonomous systems, and more precisely to unmanned airborne vehicles (UAVs) used for traffic surveillance. The first three years with focus on basic research will result in methods and system architectures to be used in UAVs. Because of the nature of the work most of the testing is being made using simulated UAVs in simulated environments, even though real image data has been used to test the vision module. In a second phase of the project, however, the testing will be made using real UAVs.

The WITAS project is a research cooperation between four groups at Linköping University. More information about the project can be found at [14].

1.2 General system architecture

The general architecture of the system is a standard three-layered agent architecture consisting of

- a deliberative layer mainly concerned with planning and monitoring,
- a reactive layer that performs situation-driven task execution, and
- a process layer for image processing and flight control.

Of particular interest for this presentation is the interaction between the reactive layer (currently using RAPS [5] [6]) and the process layer. This is done in terms of *skills*, which are groups of reconfigurable control processes that can be activated and deactivated from the reactive layer, and *events* that are signals from the process layer to the reactive layer. Events can both carry sensor data

and status information. In the rest of the paper, we will refer to the deliberative and reactive layers as the decision-making module.

Besides vision, the sensors and knowledge sources of the system include:

- a global positioning system (GPS) that gives the position of the vehicle,
- a geographical information system (GIS) covering the relevant area of operation, and
- standard sensors for speed, heading and altitude.

Currently, the system exists as a prototype implementation operating in a simulated environment, and some functionalities, e.g., GIS and deliberation, only exist in simplified forms.

1.3 Related Work

The areas in which most work has been produced with relevance to the issues presented in this document are event/episode recognition and active vision.

Pioneering work in the event/episode recognition has been done by Nagel [11] and Neumann [12]. The aim of their work was to extract conceptual descriptions from image sequences and to express them in a natural language. As the focus of the work is on the natural language aspect, all vision processing up to a complete recovery of the scene geometry including classified objects was done by humans.

The Esprit project VIEWS by Buxton, Howarth and Gong is one of the most interesting works on episode recognition in the traffic domain [8][2][9]. In this work, video sequences of the traffic flow in a roundabout are examined and events such as overtaking and cars following each other are recognized. A stationary and precalibrated camera is used, and the system presupposes an intermediate-level image processing that detects moving objects and estimates various properties of these objects. Given this information, and the ground-plane representation, the system can recognize simple events, e.g., a car turning left, and episodes, e.g., a car overtaking another car, which are composed of simple events using a Bayesian belief network. Focus of attention and deictic pointers are used to increase the performance of the system.

Active or *animate vision* is currently an active area of research in computer vision. One of the pioneers of this area is Ballard [1] who has pointed out that vision is an active process that implies gaze control and attentional mechanisms. In contrast to traditional computer vision, active vision implies that the tasks direct the visual processing and establish which parts of the image are of interest and which features should be computed. By reducing the complexity and accelerating scene understanding, active vision opens up the possibility of constructing continuously operating real-time vision systems. Our approach is fully within the active vision paradigm since the executing tasks at the decision-making level select what part of the image the vision module processes and what features are computed. Deictic pointers are also created to objects of interest and the vision module is focused on these objects.

Our aim is to create an integrated vision and decision-making component capable of complex behaviors. This was a goal also for the Esprit project VISION

As Process [3]. It integrated a stereo camera head mounted on a mobile robot, dedicated computer boards for real-time image acquisition and processing, and a distributed image description system, including independent modules for 2D tracking and description, 3D reconstruction, object recognition, and control. This project has similarity with our project even if the application domain is different. In particular, both projects include active vision, focus of attention, scene manipulation and the need of real-time performance. We intend to use some of the methods developed during the Vision As Process project and reconsider them in the context of our application.

Reece and Shafer [13] have investigated how active vision can be used for driving an autonomous vehicle in traffic. They address techniques for requesting sensing of objects relevant for action choice, decision-making about the effect of uncertainty in input data, and using domain knowledge to reason about how dynamic objects will move or change over time. Autonomous vehicles have been investigated also by Dickmanns [4].

A project for autonomous take-off and landing of an aircraft is currently under development by Dickmanns [7]. Conventional aircraft sensors are combined with data taken from a camera mounted on a pan and tilt platform. The camera data is mainly used during the final landing approach to detect landmarks and possible obstacles on the runway. Regarding vision, this work is mainly focused on object recognition.

The RAPS system used in our reactive layer has been employed previously to control a vision module [6]. Similar to our approach, the executing tasks call visual routines that execute specific image processing routines. The added difficulty in our case lies in the fact that the anchoring between symbolic and visual information is complicated by the dynamics of the objects in the scene. Anchoring between symbolic and perceptual information has been considered in the Saphira architecture [10], but also in this case mainly for static objects.

To summarize, the aspects that are more extensively studied in the above projects are event/behavior recognition, active selection of vision processing algorithms, and focus of attention. Not so widely explored are general methods for integration of static and dynamic knowledge, continuous support of the vision module by the decision-making module on the basis of short term prediction, and general methods for anchoring of symbolic to visual information in dynamic scenes.

2 Integration of vision and decision-making systems

In this section we discuss several important issues related to the integration between the vision module and the decision-making module. As a result of this discussion we propose the intermediate structure called Scene Information Manager, elaborated in the next section.

2.1 From image domain to symbolic information

The data required by the decision-making module is mainly about the road network and about moving objects and their position with respect to the road network. For example, if the airborne vehicle is pursuing a car, it needs to know in which road the car is, where along the road it is, and in which direction the car is moving (dynamic information). It also needs to predict future actions of the car based on the structure of the road network (static information). Typically, the static information is retrieved from a GIS, and the dynamic information is produced by the vision module.

The integration of static and dynamic information can be done in several ways, but the solution implies in general that symbolic data, e.g., the label of the road on which the car is traveling, has to be accessed by means of information derived from the image domain, e.g., image coordinates of the car. This task depends on low-level parameters from the camera calibration and, therefore, does not fit the abstraction level of the decision-making module. However, to access the static information image coordinates have to be transformed into some absolute reference system, using the information in the GIS. Database access is not a typical image processing task, and therefore the solution does not fit the abstraction level of the image processing module.

2.2 From symbolic information to image domain

The above description also applies to the information flow which goes from the decision-making module to the vision module. For example, if the decision-making module decides to focus its attention on a specific car (which can be outside the current field of view), the knowledge about this car is represented in symbolic form, e.g., that the car is located at a certain distance from an end point of a specific road. To solve this task, however, the vision module must know the angles by which the camera has to be rotated in order to point the camera at the car. Hence, there is a need for translating symbolic information (road/position) to absolute coordinates from which the camera angles can be derived given the absolute position of the UAV.

2.3 Support and guidance of visual skills

Knowledge about the road network should help the vision processing and give hints as to what the vision module is expected to find in the image. For example, knowledge about roads and landmarks that are expected to be found in the image can greatly facilitate the vision module in recognizing objects in the image that correspond to road network elements. Knowledge about the road network structure and its environment can also avoid failures in the image processing. For example, if the vision module is tracking a car and the car disappears under a bridge or behind a building, the vision module can get confused. However, this situation can be avoided by giving information to the vision module about the

presence of the occluding objects and the coordinates of the next position where the car is expected to reappear.

The basic mechanism of this support is prediction. Prediction, e.g. of future positions of a car, or of whether the car will be occluded by another object, is usually a high-level processing which relies on an understanding of the concepts of cars, roads, and occlusions. On the other hand, the final result of this prediction will be used directly in the low-level parts of the vision module. This implies that the prediction processing has to be made by some type of decision-making and image processing hybrid.

2.4 Support of decision making

The vision system delivers information in the same rate as camera images are processed, and on a level of detail which is not always relevant to the decision-making module. Thus, there is often a need to filter and compile information from the vision module before it is presented to the decision-making module. For instance, some vision skills compute uncertainty measures continuously, but these measures are only relevant to decision-making when they pass some threshold.

2.5 Discussion

From the above presentation we can conclude that by employing an intermediate structure, located between the high-level decision-making module and the low-level vision module, some important issues related to the integration between the two module can be handled. This structure is dedicated to translating symbolic references, e.g., in terms of labels, to either absolute or image coordinates, and vice versa. To do so it needs access to a GIS in order to retrieve information about the road network, both in terms of the connection topology and the shapes and positions of each road segment. By means of this information it can make links between static information (the roads) and dynamic information (the cars). In order to translate between absolute world coordinates and image coordinates it needs to have access to a reliable positioning system which continuously measures the position and orientation of the UAV and its image sensors, e.g., using GPS and inertial navigation.

Using the information which it stores, this structure can provide support to the vision module based on high-level prediction of events such as occlusion. It can also act as a filter and condense the high frequent low-level information produced by the vision module into low frequent high-level information which is sent to the decision-making module.

The intermediate structure proposed above is here called the Scene Information Manager (SIM), and it is presented in the following section.

3 The Scene Information Manager

Given the complex interaction between vision processing and decision-making, it is apparent that there is a need for a structure that can store static and

dynamic information required, and that also satisfies the needs of vision and decision-making as described in the previous section. The Scene Information Manager (SIM), figure 1, is part of the reactive layer and it manages sensor resources: it receives requests for services from RAPS, in general requests for specific types of information, it invokes skills and configurations of skills [1] in the vision module (and other sensor systems), and it processes and integrates the data coming from the vision module. Currently, a standard color camera is the only sensor resource present, but one can expect the presence of additional types of sensors in the future. In the following sections, we present the functionalities of the SIM.

3.1 World model and anchoring

The SIM maintains a model of the current scene under observation, including names and properties of elements in the scene, such as cars and roads, and relations between elements, e.g., a car is in a position on a specific road, or one car is behind another car. What is stored is mainly the result of task-specific service requests from the decision-making levels, which implies that the model is partial;

[1] A configuration of skills is a parallel and/or sequential execution of skills.

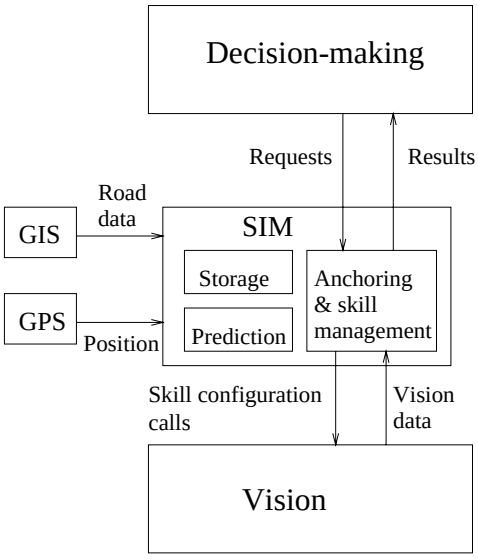

Fig. 1. Overview of the Scene Information Manager and its interactions with decision-making and vision.

with some exceptions, information that has not been requested is not registered. The SIM also maintains a correlation between the symbolic elements and image elements (points, regions). This correlation (anchoring) is done in terms of shape and color information, and reference systems which are independent of the position and orientation of the camera. For instance, if a service request refers to a specific car by its name, the SIM looks up its coordinates and signature and provides these as parameters to the vision module. The vision module is then responsible for performing the processing required to find the car in the actual image. Likewise, the SIM is also capable of finding the symbolic name of an object given its position in the image, and of assigning names to objects that are observed for the first time.

Finally, the SIM contains mappings from symbolic concepts to visual representations and vice versa. For instance, colors, e.g., "red", are translated to color data, e.g, RGB values, and car models, e.g., "Mercedes", can be translated to geometrical descriptions.

3.2 Skill management

The SIM is responsible for managing service requests from the decision-making levels such as looking for a car with a certain signature and calling the appropriate configuration of skills with the appropriate parameters. These parameters can include cars, which are are denoted with symbolic names in the request and translated ("de-anchored") when passed on to vision routines, and concepts, which go through appropriate translations. The SIM is also responsible for returning the results produced by the vision skills to the task that requested the service, and to update its information about the relevant objects. In order to do this, it has to keep track of the identities of relevant objects. For instance, if a service is active for tracking a specific car, then the SIM must maintain information about what car is being tracked (indexical referencing).

Furthermore, skill management involves combining the results of different visual processes, or adapting or compiling the output of visual processes to a form which is more suitable for decision making. In particular, it involves reducing the amount of information sent to the decision-making module by detecting and notifying when certain events occur, such as when a given threshold is passed. For instance, the visual data include certainty estimates, and the SIM determines whether to notify the decision-making module that the certainty is too low or to confirm that it is good enough. This treatment of uncertainty supports making decisions about taking steps to improve the quality of data when necessary, without burdening the decision-making module with continuous and detailed information about measurement uncertainties.

3.3 Identification of roads

The information stored in the SIM is mainly the result of active skills; objects that are not in the focus of some skills will simply not be registered. The only

"skill-independent" processing going on is the identification of roads and crossings, based on information about the positions and geometries of roads extracted from the GIS. This information is used to find the parts of the image corresponding to specific roads, which enables determining the position of cars relative to the roads. This is the most important example of integration of static and dynamic knowledge in the system.

This functionality can be implemented in several ways, and two quite different approaches have been tested. One is based on tracking landmarks with known world coordinates and well-defined shapes which are easily identified in an aerial image. From the world coordinates and the corresponding image coordinates of all landmarks, a global transformation from image to world coordinates (and vice versa) can be estimated assuming that the ground patch which is covered by the image is sufficiently flat. A second approach uses the shape information about each static object, e.g., roads, and measurements of the position and orientation of the UAV's camera to generate a "virtual" image. This image "looks" the same as the proper image produced by the camera, but instead of intensity values each pixel contains symbolic information, e.g., road names, position along the road, etc. The virtual image works as a look-up table which is indexed by image coordinates.

Since it relies on tracking of several landmarks, the first approach is more robust but less effective and versatile than the second approach which, on the other hand, is less robust since it depends on having enough accurate measurements of the camera's position and orientation.

3.4 Prediction

The information stored in the SIM is not just what is obtained from the most recently processed image, but includes the near past and a somewhat larger region than the one currently in the image. Past information such as the position of a car two seconds ago, are extrapolated to find the car again after being temporally out of focus, thereby increasing the robustness of the system and extending its functionality. Such extrapolation might involve formulating alternative hypotheses, like a number of possible positions of the car. In this case, vision is directed to check one hypothesis after the other until either the presence of the car is confirmed in one of the positions or there are no more hypotheses to consider. Likewise, prediction can also aid in determining if a newly observed car is identical to one observed a short time ago.

3.5 Conclusions

In conclusion, the SIM has a number of functions, ranging from storage and parameter translation to supportive prediction. In addition, the SIM provides a flexible interface between the vision system and the decision-making levels, which supports modifying concept definitions and exchanging visual processing techniques with preserved modularity. In the next section we present two specific

tasks implemented in our system, namely looking for and tracking a car of a specific signature.

4 Examples

In this section we illustrate the most significant points of the current implementation of our system by means of some examples which include looking for cars of a specific signature and tracking a car. For the case of tracking a car, it is shown how the vision module can be supported during the tracking procedure with high-level information regarding occlusion. The examples have been simulated on an SGI machine using MultiGen and Vega software for 3D modelling and animation.

4.1 Looking for and tracking a car

The goal of the UAV is here to look for a red Mercedes that is supposed to be near a specific crossing and, once found, follow the car and track it with the camera. During the simulation sequence, the UAV flies to the crossing and, once there, the decision-making module requests the SIM to look for the car. Rather than sending the service request directly to the vision module, it is first processed by the SIM which invokes a skill configuration and translates the symbolic parameters of the skills to what they mean in vision terms. In this case, color values, e.g., RGB-values, are substituted for the symbolic name of the color, and the width and length of the car are substituted for the name of the car model. Furthermore, the absolute coordinate of the crossing is substituted for its symbolic name. The vision module then directs the camera to that point, and reports all cars its finds which fit the given signature within a certain degree of uncertainty. In this particular case, two cars are found, see figure 2, and their visual signature (color/shape) together with their image position are sent to the SIM. Here, image coordinates are translated into symbolic names of roads and positions along these roads. For each of the two cars, a record is created in the memory part of the SIM, and each record is linked to the corresponding road segment already present in the memory. These records also contain information about the actual shape and color which can be used later, e.g., for reidentification. Once established, the linkage between cars and roads can be used by the decision-making module for high-level reasoning.

So far, most of the activities has been taking place in the SIM and in the vision module. However, since more than one car has been reported to fit the description, the decision-making module has to decide on which of the two cars it will follow or, alternatively, to make more measurements in order to obtain a better support for its actions. In this case, it chooses to follow one of the two cars, and it requests that the chosen car is to be reidentified (since it may have moved since last seen) and then tracked. However, since the decision-making module only has a symbolic references to the car, the SIM must translate this reference to a world coordinate which the vision module, in turn, can translate

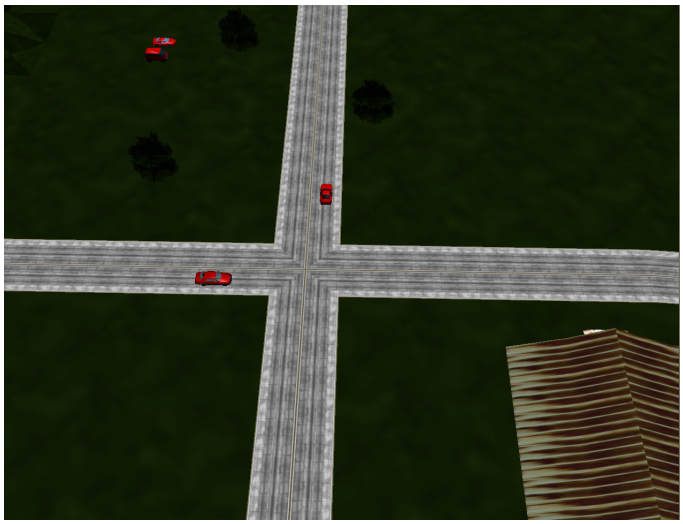

Fig. 2. Two cars of the right shape and color are found at the crossing.

Fig. 3. The camera is zooming in on one of the two cars in the previous image.

into camera angles. The SIM also provides the previously measured signature (color/shape) of the specific car to be reidentified. Assuming that the car is sufficiently close to its latest position it can now be reidentified and the tracking can start. Figure 3 shows the situation where the camera is zooming in on the car just prior to starting the tracking. In the case when there are ambiguities about which car to track, or if none can be found, the vision module reports this directly to the decision-making module.

If the chosen car turns out to be wrong, e.g., an integration of measurements shows that it does not match the shape of a Mercedes, this is reported back to the decision-making module. In this example, there is one more car which fits the given signature, but it is out of sight from the camera and some time has elapsed from when it was last seen. However, using the information stored in the memory of the SIM, the prediction component of the SIM can make high-level predictions on the whereabouts of the second car. Consequently, upon receiving a report on the tracking failure of the first car, the decision-making module sends a request to reidentify and track the second car, this time based on high-level prediction of the car's current position. This is solved in the same way as for the first car, with the exception that there are several options regarding the car's position, since it was approaching a crossing when last seen. Each of the positions is tested using the reidentification skill of the vision module until a matching car is found or none can be found.

It should be mentioned that during the tracking operation, the vision module is not just responsible for the camera motion. The image position of the car is constantly being measured and then translated by means of the SIM into symbolic links relative to the road network.

4.2 High-level support during tracking

In the second example we illustrate another use of the prediction component of the SIM. During tracking the prediction component uses information about the tracked car and of the surrounding road network to perform high-level predictions about the position of the car, as was mentioned above. However, the prediction functionality is also used in case of occlusion of the car, e.g., by large buildings or bridges.

The prediction module regularly estimates the expected position of the car and, in case the car disappears or there is a significant difference between the expected and measured positions, it checks the presence of occluding objects by consulting the geographical information system. If the presence of an occluding object (e.g. bridge or tunnel) is confirmed, the vision module receives information about where the object is going to reappear. If there is no record of an occluding object, the vision module uses predicted information about the car's position for a pre-specified amount of time. If the car reappears, normal tracking is continued, and if it does not reappear, a failure message is sent to the decision-making module.

Figure 4 shows two identical cars, one traveling along a road which makes a bend under a bridge, and one which travels on the bridge. In this example,

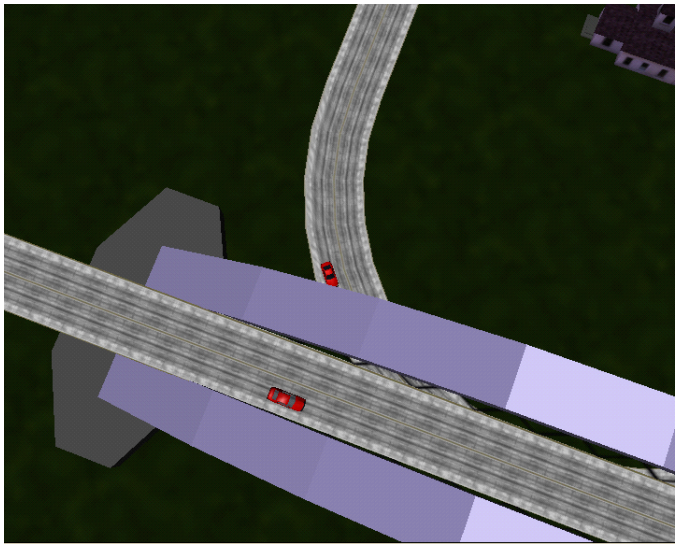

Fig. 4. The UAV is tracking a car which soon will disappear under the bridge. A second car is on the bridge.

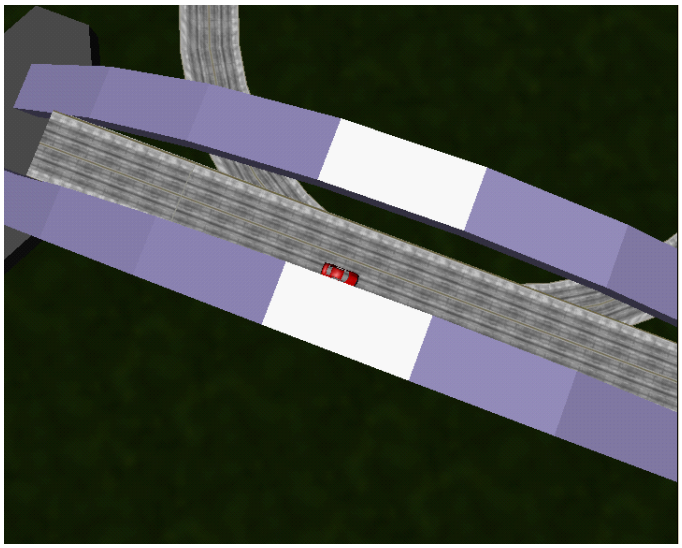

Fig. 5. The tracked car is occluded by the bridge and is therefore likely to be confused with the second car.

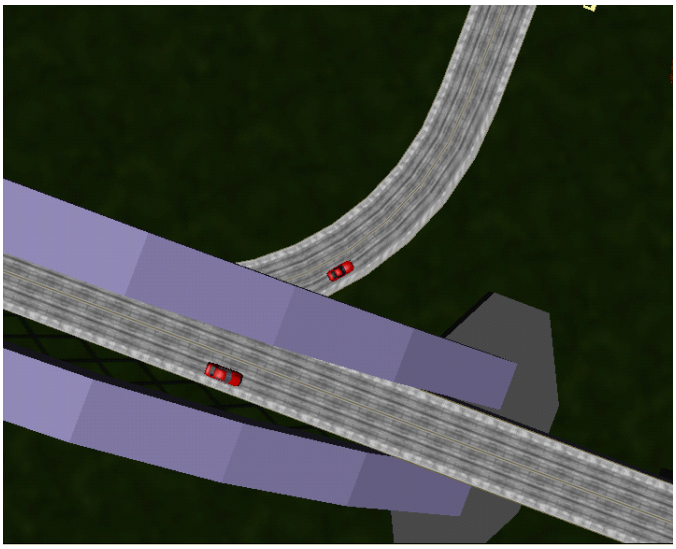

Fig. 6. The car reappears from under the bridge and can therefore be tracked again.

the UAV is tracking the first car which soon will disappear under the bride and, even worse, a few moments later the second car will be a position in the image where the first car would have been, had it not be occluded, figure 5. Using the high-level prediction provided by the SIM, the vision module reidentifies the correct car when it reappears from under the bridge, figure 6.

5 Conclusion

This paper presents a system which integrates vision and decision-making in an autonomous UAV for traffic surveillance. A number of issues related to the integration have been considered: integration of a priori and dynamically acquired knowledge, anchoring of symbolic data into visual data, focus of attention, supporting the execution of visual skills by decision-making, and handling of uncertainty.

A structure, the Scene Information Manager, where these issues are addressed has been presented. Integration of a priori knowledge and dynamically acquired knowledge is achieved by continuously matching elements in the GIS and corresponding elements in the current image. Anchoring is performed during service requests and transmission of visual data to the decision-making module, using stored data and short term prediction. Certainty of visual data is reported to the SIM where it is elaborated and transmitted when necessary to the decision-making module. Uncertainty due to extrapolation of data is handled by the SIM by guiding the vision in checking one hypothesis after the other. Finally, short term prediction is used to support the vision module and anticipate failures.

6 Acknowledgment

The simulation data presented here is the result of a teamwork in which participated, besides the authors, Johan Wiklund, Thord Andersson, Gunnar Farnebäck, Tommy Persson, and John Olsson. The WITAS project is fully sponsored by the Knut and Alice Wallenberg Foundation.

References

1. D. H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–87, 1991.
2. H. Buxton and Shaogang Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.
3. J. L. Crowley and H. I. Christensen, editors. *Vision as Process*, Esprit Basic Research Series, Berlin, 1995. Sringer-Verlag.
4. E. D. Dickmanns. Vehicles Capable of Dynamic Vision. *IJCAI-97*, Nagoya, Japan, 1997.
5. J. Firby. Task Networks for Controlling Continuous Processes. Proceedings of the Second International Conference on AI Planning Systems, 1994.
6. J. Firby. The RAP language manual. Technical Report AAP-6, University of Chicago, 1995.
7. S. Fürst, S. Werner, D. Dickmanns, and E. D. Dickmanns. Landmark navigation and autonomous landing approach with obstacle detection for aircraft. In *AereoSense97*. Orlando, FL, 1997.
8. R. Howarth. *Spatial representation, reasoning and control for a surveillance system*. PhD thesis, Queen Mary and Westfield College, 1994.
9. R. Howarth. Interpreting a dynamic and uncertain world: task-based control. *Artificial Intelligence*, 100:5–85, 1998.
10. K. Konolige, K. L. Myers, E. H. Ruspini, and A. Saffiotti. The Saphira architecture: A design for autonomy. *Journal of Experimental and Theoretical Artificial Intelligence*, 9(1):215–235, 1997.
11. H. H. Nagel. From image sequences towards conceptual descriptions. *Image and vision computing*, 6:59–74, 1988.
12. B. Neumann. Natural language description of time varying scenes. In *Semantic Structures*, pages 167–206. Lawrence Erlbaum associates, Hillsdale, NJ, 1989.
13. D. A. Reece and S. A. Shafer. Control of perceptual attention in robot driving. *Artificial Intelligence*, 78:397–430, 1995.
14. `http://www.ida.liu.se/ext/witas/eng.html`