Interpreting a dynamic and uncertain world: high-level vision

Richard Howarth

Department of Computer Science, Queen Mary and Westfield College,
Mile End Road, London E1 4NS, UK.
howarth@dcs.qmw.ac.uk

Abstract

When interpreting a dynamic and uncertain world it is important to have a high-level vision component that can guide the reasoning of the whole vision system. This guidance is provided by an attentional mechanism that exploits knowledge of the specific problem being solved. Here we survey work relevant to the development of such an attentional mechanism, using surveillance as an application domain to tie together issues of spatial representation, events, behaviour, control and planning. The paper culminates in a brief description of HIVIS-WATCHER a program that makes use of all these areas.

Keywords: High-level vision, spatio-temporal representation and reasoning, behaviour, deictic representation, probabilistic reasoning.

Submitted to Artificial Intelligence Review September 12, 1994.
1 Introduction

Now that robust low- and intermediate-level vision systems are being developed (see for example, Murray and Buxton [72], Murray et al. [73], Worrall [112]) it is worth considering how these vision systems can be combined with AI techniques to form systems that perform more high-level, cognitive processing. In this paper we reduce the extent of the survey by considering the less complex vision problem of surveillance performed by a single static agent (a fixed camera) which we call the "official-observer", and which has the objective of understanding the activity of other participants in a known wide-area scene. We call this restricted problem the surveillance problem. In the modern world there is an increasing use of surveillance, resulting in the need for automatic or semi-automatic methods for processing the dynamic input data. Surveillance concerns more than just observation, in addition to having some intelligence and knowledge, the perception performed is much more purposive, complying to some known task.

To perform surveillance we need to reason about the activities of the objects that are perceived. The process by which this visual perception is performed is complex and will not be fully addressed here. However, since perception is an important part of surveillance we need to sketch its relationship to this survey. Figure 1 shows how we can separate vision into three stages: low-level (or early), intermediate-level and high-level (or late). Low-level vision is the best understood (Horn [49], Marr [65]), and concerns visual receptors, be they from a television camera (basically just producing a 2D array of grey level intensity values) or biological, with low-level processing using visual primitives that act on the results from the visual receptors, to provide image features such as edges, corners and flow vectors. Intermediate-level vision is less well understood, and concerns the recognition of objects (e.g., model matching and tracking). Marr describes a framework for 3D interpretation, that begins to address intermediate-level vision. High-level vision is the least well understood and concerns the interpretation of the evolving information that is provided by intermediate-level vision as well as directing what intermediate-level visual processing should be performed. In progressing up these levels we see that image oriented information is at the lower levels and the more abstract, symbolic descriptions are at the higher levels. It is the development of high-level visual processing that allows the results from intermediate-level visual processing to be used for reasoning over longer time scales and we can use this to obtain a greater understanding of what is going on in the field-of-view. In this survey we concentrate on the role of high-level vision, with emphasis placed on how what we know about an environment affects its interpretation.
1.1 Basic architecture

By using a single fixed camera we ignore difficult issues associated with active cameras and multisensor fusion, while providing ample visual data to address surveillance tasks. Figure 1 shows where the interface falls between intermediate- and high-level processing. We choose the separation as an initial framework which mirrors how computer vision and AI are generally treated (for example, Charniak and McDermott [19], Corrall and Hill [22]) and because drawing the interface in this position allows the results from intermediate-level processing to be collected, as a stream of compact encodings for subsequent high-level processing. Later on we briefly discuss the problems caused by this interface.

1.2 The official-observer

In the surveillance problem the official-observer is not a partner in the interaction taking place in the scene, and although the official-observer visually attends to them, the observed people/objects\(^1\) are not necessarily aware of this. The things that are relevant to the official-observer are likely to differ greatly from those of the parties involved in the interaction. This permits the official-observer to see at the same time more and less than what is seen by the participants, with only the fragments of action manifested by them being accessible for observation. To understand these fragments, the official-observer has to use knowledge of similar patterns of interaction with the goal of constructing the motives of the actors from that part of its knowledge pertinent to the observation. The knowledge used by the official-observer is therefore different to that used by the participants in the interaction. Even with this difference there is a chance, sufficient for most practical purposes, that the subjective meaning of the actor’s acts can be understood. This chance increases with the degree of anonymity and standardisation present in the observed behaviour.

1.3 Issues

To address the surveillance problem requires that the official-observer maintain a model of the static scene\(^2\) and dynamic object property values in its world. This model should roughly consist of:

- some theory of how the domain operates: rules governing typical object behaviour and routine observation tasks, assumptions about the set of identified events, the truth of propositions, and so on;
- data from the sensor (results from intermediate-level visual processing), a priori knowledge about the observed scene, and a specification of the surveillance task or policy.

The survey separates the elements of this model into three methodologies: spatial representation, events and behaviour, control and planning. It is by using these additional forms of knowledge about the task and scene that it becomes feasible to understand processed image data. The progression through the survey corresponds to an investigation of how spatial representation and then reasoning about the data contained in that representation, can best be performed in the context of the surveillance problem. Issues connected with spatial reasoning are addressed in relation to understanding the activity of agents in an environment, and it is consideration of these issues about events and behaviour that show the necessity of control and planning.

In this survey we first describe work relevant to each of the methodologies and then describe an example program, called HIVIS-WATCHER, to illustrate how these strands can be pulled together.

---

\(^1\)We use the word “object” to refer to physical entities like cars, trams, and planes, where the person operating the machine may not be visible to the official-observer.

\(^2\)We make the simplifying assumption of reasoning in the ground plane.
2 Spatial representation

This part of the survey briefly identifies a representation that supports the various scene properties (see Howarth [50] for a longer version). Research in the field of spatial representation is extensive, although associated with other objectives such as: geographic information systems (Laurini and Thompson [61], Samet [91]), robotics and motion planning (Latombe [60], Kapur and Mundy [54]), graphics (Hoffmann [48]), qualitative reasoning (see the collection of papers edited by Weld and de Kleer [107]), and intermediate-level vision (Murray and Buxton [72], Nagel [74], and Worrall et al. [112]). Underlying these is a common representation that describes how space is structured. A good candidate for this common representation is Fleck's [32] topological approach to representing digitised spaces for both edge detection and stereo matching. In Howarth [50, 51] this use of cellular topology\(^3\) is extended to support the topological and metrical relationships identified as being required in the surveillance problem. These requirements arose from the objective of implementing a database to hold the a priori knowledge about the scene which includes ground-plane geometry and the semantic information that is attached to it. This knowledge is intended to represent the general background information that someone who knows the scene would have, including things like a model of the typical behaviour exhibited by the objects that occupy this space.

3 Events and behaviour

The main issue confronted in this survey is how to interpret a stream of results from intermediate-level visual processing using a model of a dynamic and uncertain world. In this section we consider previous work used to represent and reason about events and behaviour. We first of all consider what needs to be represented under the headings of: (1) events, episodes and verbs; (2) perception of events; (3) frames of reference; (4) situatedness. Once this background has been covered, we go onto consider options that can extend our spatial ontology to include time. This discussion takes place under the headings: (1) temporal logic; (2) tracking.

3.1 Ontological considerations

3.1.1 Events, episodes and verbs

The term "event" is widely used yet has no specific definition, it provides a useful categorisation for describing everyday happenings, allowing the continuity of everyday experience to be cut up into discrete bounded temporal units. An event is often used to denote a unit of action which can be placed in a predictable order. Schank and Abelson [92] compose events into scripts to describe typical behaviour of a customer at a restaurant such as entering, going to a table, ordering, eating and paying.

Nagel [74] describes an ontology that captures the common sense notions of the terms being used, and could be used for describing the behaviour of agents. The data supplied by the intermediate-level vision component is considered to be a signal from a sensor however complicated that sensor may be. The first thing we look for in this signal are changes that differ significantly from background noise, where noise is defined as some known property associated with the sensor. A signal change is the most primitive element used by this system. An event is any change that has been given a significant predefined semantic, providing the particular change with a symbolic record in the system. To describe combinations of more than one event we use the term episode to cover both simple and complex sequences of action. The reason for this denotation is to make events our primitive unit of change associated with the visual system.

\(^3\)Cellular topology was developed by Whitehead [108], and is also described in textbooks such as Fritsch and Piccinini [34].
This provides a hierarchical layering from events to episodes, and then to scripts. This hierarchical decomposition and relationships between the behavioural elements can be used to define a grammar where events are terminal symbols in the language to be parsed. This approach could use the syntactic methods described by Fu [37], the static semantics of an attributed grammar as described by Frost [36], the island parsing described by Corrall and Hill [22], or the compositional semantics described by Woods [111], Pereira and Warren [81] and Dowty, Wall and Peters [30]. Each of these supports a bottom-up model to form conceptual descriptions.

Badler [7] is one of the first researchers to address the problems associated with extracting a natural language description that captures the activity of the moving objects present in a sequence of images. This approach has been developed further and implemented in the ALVEN system (Tsotsos [102, 104]) which uses a realistic vision system with provision for noise and occlusion. Its final output is a frame representation from which text could be generated. The motion verbs used by Badler and Tsotsos originated in work by Miller, which was later extended in joint work with Johnson-Laird [68]. This formal analysis of motion verbs uses a logico-linguistic approach together with an informal computational framework. This has provided the foundation for other work that also draws on the linguistic literature to ground prepositions in the spatial primitives described by Herkovitz [47]. Two such projects are described by Nagel [74] (NAOS by Neumann [76] and CITYTOUR by Retz-Schmidt [84]) which are related by their common use of the road traffic domain. Both NAOS and CITYTOUR are concerned with providing natural-language descriptions from sequences of images of moving objects, that allow question-answering to take place as an off-line user query process.

Describing events by using verbs is useful because we do this in our everyday lives, however, the flexibility of language may not be needed to describe perceptual events which are really pre-linguistic. Some of the problems of specifying events are described by Reynolds [85], who recounts issues found when combining categorisation schemes developed by different observers to describe the behaviour of Rhesus Monkeys. The selection of verbs depends upon the task of the observer. The decomposition used to convert observations into events and episodes depends on what is considered to be a useful conceptual unit. For example, in some situations the episode “eating” might be a useful conceptual unit, but could itself be decomposed into finer episodes that could be individually perceived should the perceiver have both the opportunity and wish to do so.

3.1.2 Perception of events

Experimental psychology has been used to address the problem of how events are perceived. Bruce and Green [14, pages 311-374] provide an overview and, here we will pick out a few notable examples that have investigated how causal relationships are perceived. The first is based on Michotte's launching experiments [67], and the second on Heider and Simmel's work on apparent behaviour [45]. Michotte has investigated the effect of varying the temporal interval between “action” and “reaction”. These launching experiments employ the movements of simulated blocks, which bring to light properties of perceptual control but are not that illustrative in terms of everyday events. A more appropriate example is provided by Thibadeau. Thibadeau [100] describes a computational tool that provides a conceptual description of Heider and Simmel's film in which three “agents”, represented by a large triangle, a small triangle and a disc, are situated in a 2D world as shown in figure 2, in which the rectangle with a moving flap is called a house. The movement of the various shapes are described in [45, 100]. In Heider and Simmel's experiments, even in the experiment where the subjects were not prompted, a large majority of the subjects perceived the activity of the agents in terms of animate beings, chiefly of persons, spontaneously attributing intentions to the moving parts of the artificial display. Thibadeau provides an account of how action perception can be performed without object and motion
perception. Action perception involves discrete motion of objects, with emphasis placed on the conditions for the beginning and ending of motions. He uses Newtson's “point-of-action-definition”, which denotes a simple and direct report from a person about when the action is seen (for details see Newtson [77]). A temporal structure is used where an action separates a temporal history into prior states and posterior states, with the point-of-action-definition occurring just after the event has happened. The results of Newtson's work are two fold, first, that people normally segment behaviour into actions, even when they are not required to do so during an experiment, and they are remarkably unaware of their segmentations. Second, that expectation strongly affects action perception because viewers must be prepared to see an action in order for them to see it. We will return to this property of perceptual control in section 4.

Other experiments on understanding behaviour are described by Newtson [77] and From [35] who have both analysed film sequences of people performing various everyday activity which has similarity to "ethology". The commonality is because they all study the behaviour of animals in their normal environment however, ethology also includes the problem of understanding non-humans (see Bekoff and Jamieson [10], Reynolds [85]).

3.1.3 Frames of reference

Figure 3 shows an example drawn from the road traffic domain. In the figure the official-observer stands at the side of the road looking at the various vehicles. The perceiver-centered frame is called "deictic"; the environmental-centered frame is called "extrinsic", and each object-centered frame is called "intrinsic" to the object concerned. In the figure the intrinsic frame of each vehicle coincides with the deictic frame of the driver of the vehicle (i.e., the front, sides and rear of the vehicle concerned). From these we can form two useful definitions:

**Definition 1** The **local-form** is representation and reasoning that uses the intrinsic frame-of-reference of a perceived object (exocentric with respect to the observer).

---

4Deixis is used in several disciplines such as linguistics (Bühler [16], Levinson [62, pages 54-96]), the social sciences (Garfinkel [38], Heritage [46]) and spatial representation (Herkovits [47, pages 156–192], Reitz-Schmidt [84]). Deixis is the use or referent of a deictic word (e.g., I, now, this, that, here) as an aspect of a communication whose interpretation depends on knowledge of the context in which it occurs.
Figure 3: The different coordinate systems of frame values used by the official-observer and the vehicles it is watching.

Figure 4: Frames of reference: (a) an observer in “canonical position”, (b) the “canonical encounter”. Adapted from Herkovits [47, pages 158–159].

Definition 2 The global-form is representation and reasoning that uses the perceiver’s frame-of-reference, which operates over the whole field-of-view (egocentric with respect to the observer).

The global-form is not a public-world since it, like the local-form, only exists to the perceiver. We are not representing a shared world in terms of each participant.

Herkovits [47] describes how a human being learns to construct a frame-of-reference starting from two basic experiences: (1) the experience of looking straight ahead with his or her body standing upright on horizontal ground (we will call this the “canonical position”), and (2) the experience of encountering another human being face-to-face (the “canonical encounter”). In Figure 4(a) we see the horizontal plane with the observer (denoted by an eye shape) looking forward. The horizontal axes are described by terms like left/right, front/back/side, before/behind. In Figure 4(b) the perceiver in effect “combines” the point of view of the person encountered (Bob in the figure) with his or her own. The front and back axes are Bob’s and point in directions opposite to those of the onlooker. However, the right and left axes can have either the same direction as the observer’s right and left, called “mirror order”, or the opposite of Bob’s right and left, called “basic order”. We can resolve this by referring to Figure 3 and identifying that mirror order is using the global-form (definition 2) and that basic order is using the local-form (definition 1). This access to the frame-of-reference of a perceived object, allows us to analyse all the changes involved or transpositions through space of such an object in terms of our own field values and experience. This provides a more natural feel because it is the system that we use in our everyday interactions with the world.

3.1.4 Situatedness

All the agents that the official-observer perceives are engaged in activity that is situated in the environment. Each agent has goals that are unknown to the official-observer, who only sees the interactions between the agents and the environment. Winograd and Flores [110, pages 27–37] describe how the interpretation of the perceiver is not neutral, it is fundamentally social. The activities of the agents are not planned out in detail, instead they are in a state of “thrownness”. “Thrownness” is Heidegger’s (for details see [110]) term for expressing the everyday experience of continual change, which is particularly apparent when interacting with other people. When interacting it is not possible to step back, reflect and plan. Suchman [98] investigates the issue of

---

5In this paper we do not consider the vertical axis and will not use terms like top/bottom, above/below, over/under, even though this information might be available to the official-observer.
plans as something that can evolve out of "situated activity" and the use of previous experience to structure future activity. This can take the form of identified routines or more abstract plans. In this form they provide a mechanism for thinking about actions, not performing the acts themselves. Representing forms of interaction is difficult because specifying rules that govern behaviour removes much of the dynamic improvisational quality of social activity. Although no one has yet fully modelled social activity, in section 5 we sketch a more situated approach based on the idea that the observed agents are indeed improvising their activity. We assume that we can use simple local models that reflect how the official-observer interprets this perceived activity.

3.2 Computational approaches

3.2.1 Temporal logic

In the surveillance problem, we do not require a full, precise temporal reasoning ability although the reasoning that is performed clearly has an important temporal component. The actions that we are reasoning about are the "primary realities" and time is really an abstraction from them. The concept of time as a periodic process is drawn from the regular repeated ticking of clocks (see, for example, Reichenbach [83, pages 115-117]). The activity of the scene objects is perceived, but time is not. The intermediate-level vision component provides a sequence of frames marking out a non-reversible stream of spatio-temporal object position updates. This provides an unnaturally discrete view of the world because the perception of physical motion is not inherently discrete. For example, if some scene object were to oscillate at a frequency some integer multiple of the frame rate then the oscillating object would appear stationary, even though it is not. The physical motions we perceive are characterised by cyclic or continuous functions that have no natural beginning or ending. For example, in the road-traffic domain, most of the vehicles travel in one continuous motion during the time that they are visible to the fixed official-observer. From this observation of physical motion, as we discussed in section 3.1.2, the observed behaviour is segmented into actions, with each having a clear beginning and end.

Here we want to briefly investigate previous temporal representations to see whether they support the discrete input data, continuous properties implicit in the data, and the denotation of actions (see McDermott [66], Shoham [94] and van Benthem [106] for further details). There are three approaches which we call "pointwise", "interval-axiom" and "cellwise". The pointwise approach is due to Taylor [99], the interval-axiom one covers most AI temporal representations and is based upon work by Allen [4], and the cellwise approach applies Fleck's cellular topology, introduced in section 2, to the time-axis. The cellwise model is attractive because we are able to represent each frame-update as a cell within a cell complex whose underlying space is the real number line \( \mathbb{R} \). This means that cellwise time provides a model of time that supports the representation of discrete frame updates and any continuous properties identified from these updates.

3.2.2 Tracking

We have now described how time can be represented, however, this does not address how we represent the spatio-temporal properties of the dynamic scene objects. We want to consider how to represent the continuous path that each object sweeps out as it travels through the field-of-view of the official-observer, and also the expectations that each object's motion generates. Some parts of this tracking process are best performed by the intermediate-level vision component that has access to the evolving results from the image data. Dealing with compact encodings from intermediate-level vision does not provide a good foundation for some of the approaches that we might consider using. This also means that we do not have to address the problem of motion correspondence (see Cox [23] for details), since this is mostly resolved before the
Figure 5: The car on the right has its field of safe travel and minimum stopping distance displayed. Note that the field changes in size, shape and organisation as the relation between the driver and the environment changes. From Gibson and Crooks [41].

compact encodings are generated. There is the possibility of using the knowledge present in the high-level component to resolve some occlusion problems, but we do not address this issue here. We can separate the tracking problem into three approaches called "spatio-temporal", "valence", and "markers".

In the spatio-temporal approach the concern is modelling the path swept out by the object. In general this is a 3D+t or 4D problem as described by Dickmanns and Mysliwetz [28] however, here we make the simplifying assumption of operating in the ground-plane. There are a number of ways of modelling the dynamic spatio-temporal data such as building a 2D+t representation from the pose-boxes, or using the object's centroids in a simplified 2D+t representation. To correct noise present in the data, maneuvering target tracking could be used. Bar-Shalom and Fortmann [8] describe a number of options like the Kalman filter, variable dimension filter and \( \alpha \beta \gamma \) filter. The result of the target tracking could then form the basis for a 2D+t representation. Alternatively, the history represented in the state and state-transition matrices may be sufficient. Although useful for tracking object maneuvers, these filter mechanisms do not in themselves provide all the necessary spatio-temporal data necessary for reasoning about the activities of the scene objects.

The valence approach is described by Gibson [40] where he discusses how spatio-temporal predictions might be formed. The idea behind this is that in some cases the outcome of an event sequence is implicit at the outset so that it is possible to foresee the development, and possibly the end, when a perceiver sees the beginning. This is not a static problem, instead it is continuously evolving. Figure 5 provides a road-traffic domain example of how vehicle paths are seen and how anticipated locomotion is modified by obstacles in advance of the actual process. Gibson describes how the totality of possible paths acts as a guide to locomotion, being the process that keeps you to the middle, as shown in the figure where the vehicle is kept in the centre of the field of safe travel. See Gibson [40, pages 223-237] for discussion on this and other optical information necessary for control of locomotion. Should we want to implement the vehicle path part of this valence approach we could use the maneuvering target tracking mechanisms to provide an estimate of the field-of-travel for an object even though these filters are poor at providing predictions (Goodwin and Sin [43] describe some better solutions). It
is certainly the case that expectation plays a key role in understanding object activity in the surveillance problem.

The marker approach does not retain any temporal history. It is closely associated with the attentional mechanisms used in early vision which have been investigated by Koch and Ullman [57] (which Chapman [17] has implemented) and Mozer [71] and Tsotsos [103] (which he has implemented with Culhane [24]). The attentional mechanism is used to tune the early visual input, selecting a small portion of the visual stimuli to process. The compact encodings provided in the surveillance problem could have been identified using an attentional approach, although the current problem description does not allow feedback for directing the early visual attentional processing. We can approximate this, using an approach described by Agre [2] and Chapman [17] where a buffer is used to store the new frame-update for each object under some form of unique identifier that allows its contents to be updated with each new frame. Here the unique identifier acts as the pointer address that can be held by a marker. No frame-update history is maintained and the changing object position values give the effect of the marker tracking the object.

3.3 Summary

Once we identify the frame of reference of the observed objects as being an important component of surveillance, we find that this needs the on-line reasoning provided by the situated approach. These combined approaches present an appealing framework in which to investigate a more dynamic approach to surveillance because they are able to capture the improvised interactions present in everyday activity. The adoption of the indexical approach brings with it the requirement of more complex control.

4 Control and planning

We discuss control issues by first considering: normal behaviour, context and attention, and deictic representation, and then go on to explore how these have been applied.

4.1 Ontological considerations

4.1.1 Normal behaviour

An important influence upon our understanding of the environment and the use to which it is put comes from our social context. This has been investigated by Garfinkel [38] whose objective is the recognition of activity in a social context. He has performed experiments [38, pages 40-44] that demonstrate the presence of "normal" behaviour or "maxims of conduct", and which Heritage [46, page 117] has summarised as:

Definition 3 Reflexive accountability exists when each of the participants hold one another accountable. This social world will exist when the following three conditions hold: (1) that the social participants are "aware" of the norm, (2) that they are, on occasion, capable of reflexive anticipation of the interpretive consequences of breaches of the norm, and (3) that they attribute the conditions (1) and (2) to each other.

It is debatable whether the official-observer is part of the social world since it does not communicate with the participants in the environment, however, knowledge of the norm is necessary to understand what is taking place in the social world. The norm becomes most apparent when it is breached (or "broken" in the sense of Heidegger [110, pages 36-37]). Heritage [46, page 116] provides an illustrative example where a greeting is not returned. This reflexive accountability of the actors keeps perceivably normal conduct "on the rails" because they are able to anticipate some of the interpretations that their exercise of the options will give rise. This has similarities with Gibson's "valence" approach [40], described in section 3.2.2. We can develop
Figure 6: The top line appears to be composed exclusively of numerals, and the bottom line of letters. However, the symbols representing 8 and 13 are common to both (see Bruner [15] for details).

Figure 7: Examples from Yarbus [113] illustrating how a given perceptive task affects selective attention. (a) "An Unexpected Visitor" by Repin, (b) give ages of the people in the room, (c) remember positions of objects and people in the room.

an agent model that is an approximation of the official-observer's model of the scene object's normal conduct as follows:

Definition 4 A typical-object-model is a collection of rules that describe the official-observer's knowledge about how objects behave.

Further details about the typical-object-model are given in Howarth [50].

4.1.1 Context and attention

Figure 6 illustrates how the same data can evoke quite distinct interpretations and how expectation due to context, built on past and concurrent interpretation, affects interpretation of the environment. We can also change interpretation by giving an instruction, which is illustrated in Figure 7 with two examples from the large number of experiments reported in Yarbus [113]. In the figure, parts (b) and (c) show that the scanning patterns of saccadic eye movements are highly selective for the particular task at hand. This special behaviour is due to the human visual system having a small fovea, and its purpose is to quickly move this highly sensitive visual receptive area to different spatial targets. This illustrates the observation, reported in psychology textbooks such as Rock [87], that people do not scan a whole scene with the same intensity but rather, after they orient themselves, focus visually on those objects needing attention. The result of this is that we do not see everything, and miss out those things which can be ignored for the task at hand.

The problem of selecting important figures is reviewed by LaBerge [58], and described by both Rock [87] and Treisman [101] who propose attributing visual processing to two stages called "preattentive" (or peripheral) and "attentive" (or foveal). Pylyshyn and Storm [82] also uses this separation in their FINST theory by separating multtarget visual tracking into two stages, one a parallel preattentive indexing stage and the other a serial checking stage invoked in selecting a response. At the first stage simple features are preattentively registered, by what
we will call “simple operators”, which describe global features but not fine detail. Treisman provides examples of texture segregation (a prerequisite for figure-ground separation). Murray et al. [73] discuss other cue like processes such as one to detect “looming motion”.

At the second stage objects are identified using the candidates set up by the preattentive stage. Mahoney and Ullman [64] describe how the results from demonstrations of the preattentive stage support their use for directly indexing local features as long as the figure of interest is distinguished from irrelevant figures by a single one of these features. In this way preattentive processing can propose figures for use by attentive processing. This ties in with the description of deictic representation given in section 4.1.3. For example, once the first stage has identified a contiguous blob of space we can mark this location and bring specialised processing to bear on the target. This may involve things like working out what it is, by first attending to the whole object then adjusting downward to align with parts of the object.

In the first stage some of the object interrelationships that we wish to describe are similar to the early discoveries made in Gestalt psychology (see, for example, Gordon [44, pages 46–75]) concerning grouping properties, such as: proximity, similarity, good configuration, common fate (spatial and/or temporal continuity), closure, and symmetry.

4.1.3 Deictic representation

We will use the terms “entities” and “aspects” to describe the deictic representation used here. Both of these terms are taken from the work of Agre and Chapman [3, 17].

Definition 5 An entity is something that is in a particular relationship to the agent.

Definition 6 An aspect describes a property of an entity in terms of the agent's purpose.

Aspects provide information about the activities of an object, not its events. For example, the-cup-I-am-drinking-from is the name of an entity and the-cup-I-am-drinking-from-is-hot is the name of an aspect of it. Each time we obtain an aspect (say by picking up the-cup-I-am-drinking-from) its value may be different. We still have the problem of identifying events from the temporal sequence of aspect values (such as the event when the-cup-I-am-drinking-from-is-hot becomes false) but this is made more complex because the entity is not always the same (for example, during a typical day, on different occasions, I drink from a number of different cups). However there is some local temporal continuity.

An advantage of using a deictic representation is that it allows a propositional theory to be developed that is proportional to the number of properties of interest, as opposed to the number of propositional objects in the world. This means that when performing surveillance we do not need to provide a unique name for every propositional object that will ever pass through our field of view. Instead we can define a fixed, smaller number of properties of interest that the official-observer uses to describe the activities of the scene objects.

Temporal continuity can be maintained by using a marker (see section 3.2.2) to point at the scene object (entity). A similar form of indexical reference, called FINST, is described by Pylyshyn and Storm [82], who present psychophysical evidence that there is a numerical limit of four or five on the number of objects that can be tracked at one time. The ability to track four or five objects is pertinent to the surveillance problem as it can be used to place an upper bound on system complexity, since a human observer can perform surveillance with such limitations.

---

6This has been used by Agre and Chapman [8], Subramanian and Woodfill [97], and Whitehead and Ballard [109].

7Miller [69] describes similar limits on the capacity of processing information from other stimuli.
4.2 Computational approaches

4.2.1 Time Maps

The Time Map Management system (TMM) was developed by Dean and McDermott [27]. A TMM is used to efficiently reason about logical propositions whose status changes over time. In one of its simplest forms, given a set of propositions which hold at a given time, a TMM infers the propositions that may hold at a later time, after a sequence of events has occurred (see Dean and Boddy [26]). The propositions are assumed to persist in the absence of relevant change, where changes are the result of events whose effects are described by means of causal rules.

The problem with using this approach is obtaining all the necessary data and then using the TMM to perform predictive reasoning which might be more easily done without it. The TMM is likely to provide a level of consistency analysis (by continuously comparing expectations with observed activity) but at a considerable run-time cost.

4.2.2 Plan recognition

The compositional nature of the script-based approach described is section 3 lends itself to plan recognition. A review and formal treatment of plan recognition is given by Kautz in Allen et al. [5, pages 69–125], where its origins in story understanding (Schank and Abelson [92]) are described. Other approaches to this problem include Schmidt, Sridharan and Goodson's [93, 96] BELIEVER system, which interprets linear sequences of intentional statements about people in everyday environments. This has similarities to the adaptive planner of Alterman [6]. Plan recognition would be suitable for a script-based approach, and could be implemented using a TMM (see Dousson et al. [29] for example), and the uncertainty present from visual input could be addressed by using the Bayesian approach described by Charniak and Goldman [18]. However, this suitability for off-line reasoning reduces its suitability for on-line control. The approaches are better at explaining what happened than being used to understand what is happening in the here-and-now.

4.2.3 Multiagent

There is a partial overlap between the surveillance problem and multiagent planning (for example, the PHOENIX system of Cohen et al. [21]), in that the official-observer is reasoning about the dynamic processes that it observes in the same way that an agent is reasoning about the other agents that populate its world. However, we are not dealing with cooperative or adversarial agents, but just the observation of these agents as they carry on their everyday lives, and a key difference is that in the surveillance problem we are not able to communicate or control the observed agents.

Lansky [59] presents an example of a multiagent planning theory which uses the Group Element Model (GEM). GEM exploits causal independences by using explicitly defined constraints to synchronise plans. The process of plan synchronisation is not limited to a strategy of planning how to achieve each separate component task and then combine the results. Instead, a more general, adaptive strategy is used that can bounce back and forth between local (i.e., single-agent) and global (multiagent) contexts, adding events where necessary for purposes of synchronisation. Lansky describes how these planning loci can both overlap and be composed hierarchically. The use of local and global contexts would allow the official-observer to both reason about a single-object and see how this reasoning fitted into the global context. However, much of the supporting mechanism (see [59] for details) seems unnecessary for the surveillance problem because we are not planning the activity of the objects that we observe.
4.2.4 Embedded systems

An embedded reasoning system is one that is situated in the world and which operates effectively given the real-time constraints of its environment. Georgeff and Ingrand [39] describe an embedded reasoning system called the Procedural Reasoning System (PRS) which uses means-ends reasoning to govern future behaviour. PRS explicitly represents attitudes of belief, desire and intention, allowing them to be manipulated and reasoned about, providing complex goal-directed and reflexive behaviour. This framework would allow the official-observer to consider its own goals in addition to those of the scene-objects that it perceives. PRS comes close to solving most parts of the surveillance problem, however, it is not really situated in the way described in the section 3.1.4 and appears to be more closely associated to plan recognition.

To provide a more situated implementation we could adopt Agre's [2] RUNNING ARGUMENTS. This technique is difficult to describe because there are at least two intertwined theories at its core. The first is to do with planning, which we describe here, and the second is about the separation of program components, which we describe below in section 4.2.5. The RUNNING ARGUMENTS technique does not develop plans as such, although they do exist in the form of hardwired "action-descriptions", written in a language called MACNET (see Chapman [17] for details). These express what the program is to do given the data of the current and previous clock tick. This rule based form used to define the action-descriptions makes a comparison to the standard production rule form inevitable. The main difference is in how conflict resolution is addressed. When two or more "proposals" try and "fire" the same operator, any conflict is resolved by assigning rule precedence to the rule definitions. This approach allows a number of operators to be fired on each clock tick as opposed to the usual one per clock tick in production systems. This language is not suitable for planning, but can be used to describe plan like elements called "routines".

Rosenschein and Kaelbling [89, 90] present a more complex representation language called REX that is similar to MACNET in that it too is compiled to provide a resultant combinatorial logic circuit. The REX language is attractive because it has a basis in a formal logic that is similar to that of Moore [70], however, the implementation details provided in [90] are difficult to understand. In [89], Rosenschein describes how an additional layer of compilation can be added to enable proof correctness to be performed upon the supplied rule specification. This involves re-expressing the rules as clauses in a new language $\mathcal{L}$. The rules are now generated as a side-effect of performing a proof analysis on the clauses written in $\mathcal{L}$, with this proof analysis indicating the completeness of the specification, at the cost of an additional layer of compilation.

Even if we replaced the procedural rules in PRS by MACNET or REX, we would not change very much, and would still not have a situated approach. Next we investigate the second part of RUNNING ARGUMENTS.

4.2.5 Modularity

In our description of the official-observer we have not yet considered the effect of cognitive architecture on planning and control. Fodor [33] describes the traditional separation made in cognitive science between input/visual/peripheral systems and the central system. This view is not held by all researchers, for example, Brooks [12, 13] provides a different view that uses an orthogonal separation based on task-achieving behaviours. In Fodor's model, on the input side we have a collection of separate modules performing perceptual and motor processes, each of which are to a large part innate, localised to specific brain areas, and task- and domain-independent. Fodor argues that the central side is different, saying it is not modular, being instead a single homogeneous central system. The justification for this is that anything you know

---

can potentially be used in any cognitive task. Agre uses this split in RUNNING ARGUMENTS, with the “central-system” holding the rules and the “visual-system” holding a collection of information gathering operators.

The visual-system is based upon Ullman’s argument [105] for the integration of multiple visual operators that perform particular sorts of perceptual work such as tracking (described in section 3.2.2), representing shape properties and spatial relations. Previous work on operators also includes: Romanycia’s [88] description of a programming language that uses visual operators to compute properties and relations present in 2D images of simple geometric shapes; Mahoney and Ullman’s [64] description of low-level visual operators that operate on more complex shapes and curves to identify “image chunks”; and Chapman’s [17] description of visual operators used in a video game context. No one has used visual operators in a natural task domain with access to camera image data in the way described by Chapman (such as those which use “activation planes”\(^9\)). However, there is related work, including that on tracking, and work on “active contour models”\(^10\). Chapman’s work can be seen as an initial step towards the objective of defining a set of appropriate visual operators for use in real world application domains. On their own, these visual operators do not do much, but combined via rules held in the central-system, enable the system as a whole to respond to changes in the “world”.

The central system contains rules of the form described in section 4.2.3, which are used to select when an operator is to be used and what arguments are to be supplied to an activation. Crafting the rule-operator pairs into sequences (constructing routines) is done by making the result produced by one operator fulfill the input requirement of the next rule. However, this is not the only way a particular rule can be fulfilled, thus allowing the mechanism to react to similar situations that arise via a different route. A routine provides an abstraction for a common pattern of interaction between an agent’s central-system and visual-system. This reduces “planning what to do next” to a matter of deciding what to do “now” based upon how the world is “now”. Only the operators have access to the “world” data structures and the central-system only receives the results of the operators. This allows the central-system to use a simplified description of the world, that only needs to have the information necessary for making its action-selection. This restricted state ensures that the system can only reason about the current situation.

This separation of input and central system and the tight coupling between them provides one possible foundation for the situated approach, allowing us to address control without traditional planning or plan recognition. Although good for reasoning about the official-observer’s ongoing circumstance, in the surveillance problem we also need to reason about the moving scene objects.

4.2.6 Knowledge organisation

The organisation of knowledge about observed properties was mostly dealt with in section 3. Here we are considering this organisation with the objective of controlling the reasoning performed. The ALIVEN system (Tsotsos [102]) has similar objectives to the surveillance problem, it is passively understanding the difference in visual motion between pre- and post-operational film sequences taken of the left ventricle. Tsotsos generalises this cardiology example in [104] and considers the basic capabilities that should be present in an attentive vision system that is addressing time-varying phenomena. The ALIVEN project uses both atemporal and temporal control, which are integrated via a semantic network. This semantic network is a development of the one proposed by Badler [7]. The atemporal control includes goal-directed, data-directed

\(^9\)Activation planes are used to keep track of interesting regions of the image, as in Ullman’s [105] routine for computing containment.

\(^10\)Active contour models have been used to define outlines (Kass et al. [55], Cohen [20]) and dynamic regions distinguished by a particular visual property, e.g., texture and/or colour (Ivins and Porrill [53]).
and model-directed inference mechanisms with the claim that each compensates for the deficiencies in the others. The temporal control uses a modification of relaxation labelling, called a “temporal cooperative process”, to accumulate and integrate the dynamic information. ALVEN uses a rich search dimension to enable and distinguish search in image space from search in hypothesis space. Within these search spaces, focus-of-attention can be generated and maintained. Tsotsos [103] notes how this control framework can support attentive vision by using Ullman’s visual routines [105], providing an attentive foundation for the situated approach. However, these knowledge based approaches are typically less uniform and more complex in managing the control.

4.2.7 Uncertainty

Task knowledge needs to take account of the uncertainty present in perception. One solution is to use Bayesian inference, which is described in detail by Pearl [80] and Neapolitan [75]. Bayesian networks integrate a mechanism for inference under uncertainty with a secure Bayesian foundation. They have been applied to various research domains such as medical diagnosis (Spiegelhalter [95]), and model-based vision (Agosta [1], Levitt et al. [63]) which follow the approach described by Pearl and Neapolitan where once constructed the nodes and links do not change over time. This “structurally static” approach involves determining the graph structure of the network, and then supplying prior probabilities for the root nodes, anticipatory values for leaf nodes, and conditional probabilities for other nodes. The inference algorithm is then run for each addition or retraction of evidence from the leaf nodes. Breese [11] describes how to dynamically construct a structurally static network, an approach that has been applied to natural language understanding by Charniak and Goldman [18, 42]. Dynamic construction allows a smaller task specific network to be built, rather than a large general purpose one.

Some researchers have extended this approach by applying it to dynamic domains, where the world changes and the requirement is to reason over time (see for example Kjaerulff [56]). Applications include sensor based mobile robot navigation and target tracking by Dean et al. [25], and monitoring light beam sensor data by Nicholson and Brady [78, 79]. The dynamic nature of the domain is captured by extending the network over time, adding a new time-slice of nodes, at each clock-tick. These dynamic networks are Markovian, which constrains the state space and limits the history maintained by the network.

Rimey and Brown [86] do not use a dynamic Bayesian network, although their system is used in the dynamic domain of active vision. Their system called “TEA” uses specialised Bayesian networks together with a maximum expected utility decision rule that it runs iteratively to select the evidence gathering actions that maximise an expected utility criterion. TEA uses three different forms of specialised Bayesian networks that comprise: (1) a “part-of net” that models subpart relations, (2) an “expected area net” that models geometric relations, and (3) a set of “task nets”, one for each task that produces actions which are unique to the task.

All these applications of Bayesian networks to dynamic domains are relevant to the surveillance problem, with the emphasis of on-line reasoning being attractive for the situated approach. We essentially used dynamic Bayesian networks with embedded approach in [52].

4.3 Summary

We have covered a range of approaches to controlling the operation of a surveillance system. We are mainly considering the case where control of the system is seen as an important issue for performing a particular task. We have identified that this form of task dependence is only relevant in the situated approach. In section 3 we identified the use of markers and deictic representation as being necessary should we want to provide spatial reasoning that is related to an observed object’s intrinsic frame-of-reference. In this section we have covered some of
the control issues raised by this approach, identifying the usefulness of Ullman's visual routines [105], which has been integrated into larger control frameworks by both Agre (see section 4.2.5) and Tsotsos (see section 4.2.6). We are not just addressing attentional control. We require a framework that also allows conceptual reasoning about the behaviour of the observed object. This capability seems to be partly provided by RUNNING ARGUMENTS and PRS (described in section 4.2.4).

5 System overview

To show how the surveyed methodologies can fit together we briefly describe the HIVIS-WATCHER program. A more complete description is given in Howarth [50]. Figure 8 shows the top level loop which describes the order of execution of the various elements in Figure 9. HIVIS-WATCHER has a tightly coupled architecture where feedback is an important feature. This input/central split corresponds was described in section 4.2.5. The virtual world represents the left side of the interface shown in figure 1, although this architecture could be extended to support a more closed-loop architecture with the peripheral-system more directly accessing low-level properties. In HIVIS-WATCHER we only use compact-encodings.

HIVIS-WATCHER extends Agre and Chapman's approach with an agent model that can provide the official-observer with an interpretation of each agent. This local agent viewpoint is used to simplify and decompose the spatial relationships and possible interactions between moving objects. The global component, shown in Figure 9, is used to integrate the various local and global results. This interpretation via local viewpoints is expensive and to address this we use task-based control to select relevant scene objects. This attentional control allocates agents (processors) to the most task relevant objects and then collects the results into a coherent description of scene activity. As shown in Figure 9, orthogonal to this global-form/local-form separation is the AGENCY and KERNEL split which is similar, although not intentionally based upon Brooks' [12] task-achieving behaviour. The AGENCY contains rules and operators concerning how to use the local agent viewpoint. The KERNEL extends this functionality with a more abstract, global model of the world. We can still claim that the central-system homogeneity described by Fodor [33] is present, because the separation between global- and local-form is more to do with levels of attentional detail.

6 Conclusion

Particularly in surveillance, and also for interpretation in general, identifying abnormal behaviour from continuously improvised everyday activity is an important problem. In this survey we have identified the relevance of Garfinkel's work and how it is related (perhaps tenuously) to work by Agre and Chapman, Fodor, Gibson, Suchman, and Winograd and Flores. We also sketched out some initial investigations with HIVIS-WATCHER. An important problem, that still needs resolving, concerns how to provide a better definition of system behaviour and the behaviour of observed objects, than that currently done by the program designer. Whitehead and Ballard [109] describe initial work that may lead to a solution to this problem.

This survey has presented some steps made towards the final goal of developing vision systems which are able to understand the data from the everyday world that they process. There is still a long way to go, and here we have presented some relevant foundations and described the potential of using a situated approach to link AI and vision research.

Acknowledgement

This work has been funded by SERC under a CASE award with the GEC Marconi Research Centre.
Figure 8: Top level loop of HIVIS-WATCHER.

Figure 9: The elements of HIVIS-WATCHER.

References


