Artificial Intelligence Review (2005) 23:97–154
DOI 10.1007/s10462-004-4103-5

Spatial Models for Wide-Area Visual Surveillance: Computational Approaches and Spatial Building-Blocks

RICHARD J. HOWARTH
Department of Computer Science, Queen Mary, University of London, Mile End Road, London E1 4NS, UK (E-mail: howarth@dcs.qmul.ac.uk)

Abstract. Spatial models play a key role when interpreting a dynamic and uncertain world for a wide-area surveillance application. This paper presents two different views to illustrate the range of spatial models. First, we take a top-down look, where we survey various work relevant to the development of spatial models and how they have been used in AI applications. Then we take a more bottom-up look, starting with a promising spatial primitive to identify a useful foundation that can support visual surveillance applications.

Keywords: cellular topology, high-level computer vision, spatial reasoning, spatial representation, surveillance

1. Introduction

Wide-area visual surveillance provides a useful application domain for developing and applying computer vision techniques. However, visual surveillance does not end with object recognition, instead there is the additional problem of understanding what these objects are doing. This process of activity interpretation, or extracting conceptual descriptions from images, can make use of many AI techniques and can be broadly separated into three components:

- events – a behavioural description of what takes place in the scene over time in terms of the activities of the participants (e.g., objects and people) in the field of view,
- attention – a behavioural model of how the observer acts to fulfil a given visual task,
- environment – the static scene structure and contextual features associated with the observed scene.

This gives an impression of the range of issues covered by what we will call the *surveillance problem*. This paper focuses on a subpart of the *surveillance problem*, namely, how AI techniques can assist with representing the environment in order to enhance understanding of what is happening in the scene.

1.1. *Range of applications*

In this paper, the example application, used for illustration, places a camera by the side of a traffic junction with the objective of understanding the behaviour of the road users who pass through the field of view, as was done on the VIEWS project.[1] This is just one example from the range of application domains for visual surveillance outlined by Collins et al. (2000) in their discussion of how automated visual surveillance is becoming increasingly important to society and a major application of computer vision. Other example application domains include: the activities of people in various everyday locations (in rooms and corridors inside buildings; on city centre streets; on railway station platforms; in car parks; in outdoor areas; and at sporting events);[2] the movement of vehicles (on roads; on motorways/freeways; in airport holding-areas and taxiways).[3] These give an indication of the environments, often man-made structured spaces, that the *surveillance problem* is applied to.

1.2. *The surveillance problem*

To provide a framework for our discussion on spatial representation issues let us first consider the broader context. When we see the video from a camera we easily interpret the 3D spatial organisation and movement of any people and objects present. However, in reality, all the video displays is a 2D array of pixels some of which change colour or shade. It is the observer that has contributed the knowledge and skill that enables this ongoing, dynamic interpretation. This apparent ease of understanding the observed behaviour of others belies the complexity of this problem and its various elements. In addition, there is the general ill-posed nature of extracting visual evidence from the image data, made worse by noise and occlusion. Fortunately building in more knowledge about the task and application can help overcome some of these problems.[4] Additionally the *surveillance problem* is an instance of visual understanding where we can make some simplifying assumptions:

- We are using a known, static 2D ground-plane. This assumption is made to enable the recovery of depth information[5] and as it is a difficult problem to solve implementations may just deal with the 2D image-plane representation. Probably because there is no need to recover 3D world positions in their work. Which is a shame as this is just the kind of information that tends to be assumed to be present by AI researchers.

- We are using a single, fixed camera. This assumption is often made by surveillance implementations although it does produce a restricted view of the scene. There are an increasing number of exceptions, such as installations that use active or multiple[6] cameras or that place them on a mobile platform.[7]

These assumptions may not be necessary for all surveillance systems, but do help to sketch the *surveillance problem*. A more complete picture is given by Regazzoni et al. (2001) in their excellent historical account, and analysis of future needs, that covers communication, performance, as well as multisensor use, and by Collins et al. (2000, 2001), who recount the importance of tracking and classification and activity analysis for addressing the problem of automated surveillance.

1.3. *Using environmental context*

While calling it a shortcoming would be wrong most surveillance systems do not take advantage of the static camera's constant, ambient view of its surroundings and use knowledge of each scene's static structure and content to help with situation assessment and activity interpretation. In fact most surveillance systems initial processing step often involves removing the background "clutter" and segmenting out any foreground features so that objects of interest (typically pedestrians or vehicles) can be found.[8] In contrast, to those systems that make every effort to remove and seemingly to ignore the background, here we are examining what properties present in the background could be used to help the interpretation of the activities, events and behaviours taking place in the more important foreground.

While for some computer vision applications there is no need for *a priori* knowledge about the scene being viewed, and for others the world can be its own best model.[9] It seems likely that most surveillance applications, where a static camera is used, would benefit from knowledge about the spatial arrangement of the 3D scene, using this to place any observed behaviour in context and so also aid activity interpretation. This spatial knowledge could also be beneficial for guiding attentional processing of the image to those scene areas where things happen.[10]

This contextual information is like Gibson's (1979) affordances. The environment affords some set of typical behaviours that can take place at locations within the scene. Perhaps, to make use of this environmental context, all we need to do is acquire or learn or reconstruct these spatial properties and store them in a mental model or analogical representation of the perceived reality.

1.4. *Denoting scene properties*

Not all surveillance systems ignore the environmental context. For some denoting scene properties has an important role to play. One example is "regions of interest". These are user defined areas of the image-plane or ground-plane. Collins et al. (2001, pp. 1468–1469) describe how in their multi-camera system a region of interest can be marked on a ground-plane map and be monitored by any cameras that share this area in their field of view. These include entrances and exits – those places where new objects appear and where currently tracked objects leave the scene. For a fixed camera these could be various areas around the edge of the image, or they may be special areas in the scene due to occluding features. For example, in Intille et al.'s (1997) KIDSROOM project the doorway – or doormat area in the overhead view – gets special attention. There are many other roles regions of interest can play, for example, Ayers and Shah (2001) hand label areas of the image in an office environment: that should not be occluded; where a person's head should be when performing a given action; where the content may change; and where an object to be tracked is placed.

A second example is using environmental context to improve object tracking. With a static camera the 3D scene structure imposes invariant constraints on where objects move in the image as well as how their size and speed alters with respect to distance from the camera. Buxton and Gong (1995) used this on the VIEWS project for segmenting a detected optic flow field to identify and track moving objects. Rosin and Ellis (1991) give semantic labels – ground, fence, sky – to areas in their ground level, perimeter intrusion detection, static camera, image sequences to aid the selection of appropriate classification models – human, animal, noise. Another use of this labelling technique is for describing foreground occluding image features such as lamp posts and signs – on the VIEWS project this was used to help vehicle tracking. Intille and Bobick (1995) use a distant camera that provides a wide field of view but with low resolution, so to help track the movements of people in their scene they form a bounded "closed-world" around each tracked object and use contextual knowledge about the environment to label each object's local space this, together with contextual knowledge about constraints on behaviour and likely events in their application domain, enables most tracking ambiguities to be removed.

These examples, of how environmental context has been used by visual surveillance programs, illustrate how building in more knowledge about the task and application can enhance activity analysis. Part of this knowledge can be expressed by representing the spatial features of the environment. In this paper, we look at how this can be used to help solve the *surveillance problem*.

1.5. *Spatial representation issues*

Spatial representation has an important role to play in wide-area visual surveillance applications that require understanding and interpretation of what is going on in the scene. We can identify the key requirements as being able to describe:

- The static environment that is visible to the perceiver.
- Each moving object's spatial occupancy. This is to express the perceiver's knowledge of each object's position, its spatial extent, and the part of the environment that it occupies.
- The meaning of the different parts of the environment, including both physical properties and semantic properties. These semantic properties, although not physically present, are understood by the inhabitants of the environment and can be used by the perceiver to interpret observed behaviour.

These are unusual requirements for a spatial representation and, as shown in Figure 1, there is the option of operating in either (1) the image-plane of the perceiver (i.e., what the perceiver directly observes) or (2) a ground-plane projection which provides an overhead view that is not directly available to the perceiver but which makes viewing the results easier. These requirements are part of the *surveillance problem* introduced in Section 1.2.

When we consider the spatial forms encountered in our everyday lives, it may not be surprising – as Sloman (1985) discusses – that there is such a diverse range of spatial representations, each created to support some particular purpose and viewpoint. These spatial forms seem at one and the same time unified by a common mathematical framework and yet separated by their everyday use. We can distinguish:[11]

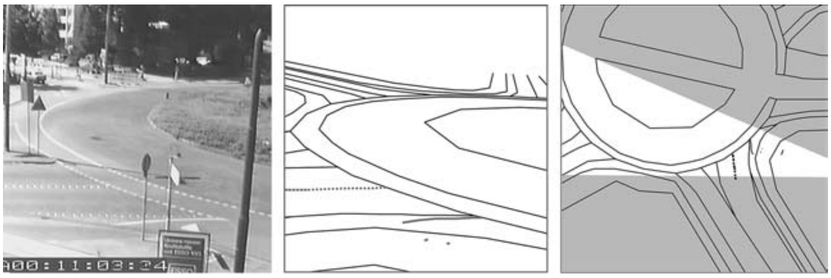


Figure 1. The scene, key image features and ground-plane projection (overhead view).

- The local space that is near us and in our field of view (*visible proximal space*), e.g., the books, desk-top and coffee mug that are in reach and visible.
- The local space that is not currently in the field of view (*occluded proximal space* and *out-of-shot proximal space*), e.g., the under side of the table, the chair the reader is sitting on, the wall behind.
- The not so local space that is in our field of view but which we can not reach (*visible space*), this space can be very large (say, the view from a mountain top) but in our everyday experience is typically smaller due to occluding objects and quality of visual resolution.
- The not so local space that we cannot see (*medium scale space*) of the building we are occupying, the corridors and other rooms we can not see but know exist.
- The *large scale space* that connects the buildings, places and locations, the routes to work, home, the shops or another town that we have visited or know about. This distinction is the difference between the current location and memory of other places. For example, Lynch (1960) describes how city dwellers remember their environment.

For a single, static camera, surveillance is situated in the *visible space* of the perceiver and we are not usually concerned with *medium scale space* or *large scale space*. This allows us to simplify our spatial representation to just those environmental features in the camera's field of view.

When we turn to consider how to represent our local spatial form in a computer, using some formal language, we find that there are constraints caused by computer storage and execution speed, which place an upper bound on what is possible. The model of the "real world" held in the machine is not complete, being just a representation of those properties that are deemed necessary for the *surveillance problem*. The representation is biased towards understanding what takes place in the scene rather than providing a realistic display on the computer screen, and can be separated into (1) the static environment, and (2) the dynamic objects that travel through the environment. With these requirements in mind let us begin to identify how they can best be fulfilled by considering related work. This investigation of various computational approaches is followed, in Section 3, with one looking at spatial building blocks – those spatial primitives that could be used to implement the approaches discussed next.

2. Computational Approaches

Here we review some related spatial representation schemes under the headings: *large scale space*, for the wider environmental context of how a space is used; *map learning*, for how the structure and use of a space can be discovered from observation; *robotics and motion planning*, for representations used to delimit usable areas of safe passage; *graphics and solid modelling*, for representations that reconstruct an environment for display purposes; *qualitative reasoning about motion and arrangements*, for representations of an environment that support reasoning about change taking place within it; *linguistic and cognitive approaches*, for denoting and communicating spatial properties; *spatial decomposition*, for techniques that make explicit the structure and organisation present in the environment; *intermediate vision*, for representations used to reconstruct what is seen in visual data of an environment; *analogical representation*, for approaches that find semantic correspondence between representations and the represented reality; and *spatial uncertainty*, to express some of the uncertainty present in the visual process and add this to the representation of the environment. Although this selection is to some extent arbitrary, it covers the main areas, each of which is likely to have something that contributes towards addressing the *surveillance problem*. In the discussion below, we will consider only a small portion of the currently available literature. For more general reviews on spatial representation in the field of AI see Chen (1990), Davis (1990), Hernández (1994), and Olivier and Gapp (1998).

2.1. *Large scale space*

While large scale spatial representations, such as Geographic Information Systems (GIS) and cognitive maps, are unlikely to be necessary for the single camera *surveillance problem*, they can provide knowledge about the larger context. Particularly if the single camera system is part of a larger network of sensors like that described by Collins et al. (2001).

Laurini and Thompson (1992) describe how GIS is more concerned with the representation of statistical and cartographic data. This data, particularly information about man-made structures, such as roads together with how flow and usage change during the day, could be of use when setting up the surveillance system and for coordinating widely separated camera sites on a road network, such as the one described by Huang and Russell (1998) and Pasula et al. (1999). However a GIS may not on its own capture all the environmental details and may not correspond exactly to the "real world". For example, Figure 2 compares a road map and an aerial photo of the same location.[12] To overcome this data correspondence problem Collins et al. (2001, pp. 1463–1464) describe how they have used geodetic coordinates (i.e., from the Global Positioning System, GPS), as their common 3D coordinate system, enabling them to use third party cartographic software and datasets to develop their site model – including maps, orthophotos, digital elevation models (DEMS), and road network graphs – and, to also incorporate airborne sensor data into their system. Illustrating how knowledge about the large scale space of a multisensor surveillance installation can be of significant assistance.

Large scale cognitive maps – as typified by the collection of papers edited by Downs and Stea (1973), and the computational models of Yeap (1988) and Kuipers (1978) – are more concerned with how we remember locations, including approaches like landmark identification (Lynch, 1960; Kuipers and Levitt, 1988) and fuzzy boundaries (Davis, 1986). These are representations of space that are intended to capture the uncertainty and distortions of memory and perception. As such, a cognitive map would make a poor site model, unless the objective is to incorporate spatial uncertainty into the representation (see Section 2.10).

This idea of constructing a memory from perceiving an environment at multiple locations overlaps with image understanding. For example, Strat and Fischler (1991) describe how their system

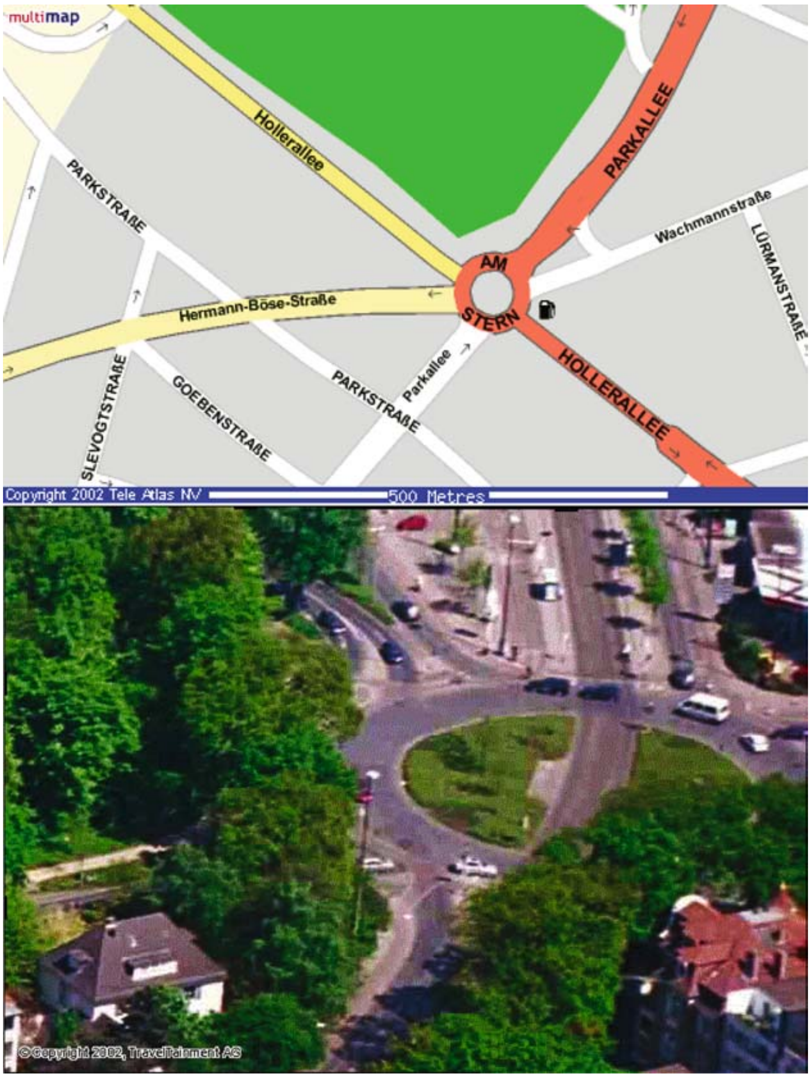


Figure 2. A map and an almost orthophoto-like overhead view of Bremer-Stern taken from an angular perspective. The map is from www.multiMAP.com and shows AM STERN, 28209, BREMEN (X:982300 m Y:6964400 m; 53:05:04N, 8:49:28E) and is Copyright 2002 Tele Atlas NV. The aerial photo is from http://www.ausderluft.de/schraeg/index.html and is Copyright 2002 by TravelTainment AG, Aachen.

CONDOR is designed to interpret 2D and 3D imagery from the Stanford Hills, California. A restricted area that contains a number of trees and bushes. CONDOR can use *a priori* knowledge, such as context derived from digital terrain models and maps together with knowledge extracted from previous interpretations, to provide contextual knowledge for interpreting the current image.[13] Building this image data incrementally up into a 3D model, in CONDOR's blackboard-like "core knowledge structure", of the location, particularly if these images are part of a sequence.

Another example is the urban scene 3D reconstruction done by Sato et al. (2002) and Mellor (2003). Sato et al. use a single hand-held camera to take a continuous video sequence from a car travelling slowly down a street and from this reconstruct a 3D model of the buildings by using these multiple views of this outdoor scene plus measured 3D positions of markers (key feature points in the image) seven of which are tracked manually in advance. In contrast Mellor's work, on the MIT City Scanning project, uses a set of digital camera images from measured "node" positions (GPS, etc.) within the scene. At each node a hemispherical mosaic of images was collected by rotating the camera about its focal point. From this collection of views around Tech Square, MIT, made under different lighting conditions due to weather and time of day, Mellor's computer program recovers dense surface patches that infer the structure of each building's facade. See Sato et al. and Mellor for discussions of related research on the 3D reconstruction of outdoor scenes.

These examples illustrate how scene reconstruction of a large scale space could contribute to site modelling, although for a single camera, on its own or as part of a multi-camera system, it is only necessary to model the scene for each individual camera's field of view. Although, sub-parts of a GIS or cognitive map can enhance local knowledge of what is present in an observed scene, there is no need for a complete large scale spatial representation of the surrounding district containing much that is extraneous to the *surveillance problem* at hand.

2.2. *Map learning*

The problem of acquiring the data for a cognitive map by exploration is related to the subject of large scale space. This concerns integrating the observed environment that is travelled through into a continuous memory structure (Davis 1986; Yeap 1988; Mataric 1991; Thrun 1998; Kuipers 2000). An example of such cognitive mapping is given by Hutchins (1983) on Micronesian navigation which contrasts the Micronesian approach with Western map use (see, for example, Monmonier 1991).[14]

More appropriate to the *surveillance problem*, than learning maps by exploration, is to learn maps from the observation of a fixed perceiver. Mohnhaupt and Neumann (1991) have addressed this problem in the road traffic domain by learning the area of space through which vehicles typically travel. To do this, they use a scene wide spatio-temporal buffer, that for each coordinate address, has a 2D array, with qualitative dimensions for speed and orientation, that accumulates a count from multiple object trajectories. Figure 3 illustrates a simple implementation of this using square cells; the results of which can be used by a prediction program, such that, given a position and an orientation, it can generate a typical path. The buffer does not represent time explicitly nor is a vehicle's extent represented. Johnson and Hogg (1996) have developed a similar model for learning a probability density function from partial trajectories. And in Johnson and Hogg (2002) they use a statistical framework of Gaussian mixtures to represent observed trajectories, formed from the base of tracked individuals, for the generation of plausible stochastic behaviour.

An alternative, to the approaches that deal with the possible movement at points in the scene, is to learn typical routes from observation. Fernyhough et al. (1996, 2000) use each object's whole shape, in the form of its image-plane silhouette and, from the motion history of these individual objects, they identify typical routes by using statistical accumulation. Figure 4 gives an illustration of how this kind of route learning operates. Fernyhough's resultant spatial paths make use of the region model developed in Howarth and Buxton (1992).

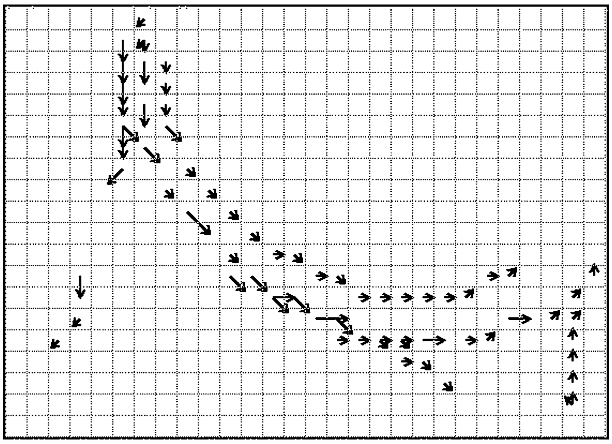


Figure 3. Results from learning the typical traffic flow on Bremer-Stern.

A different approach is taken by Makris and Ellis (2002) who model the physical extent of a route as its left and right boundary sides by accumulating normalised trajectories from tracked object centroids. Each route is stored as a sequence of nodes, where each node has, in addition to its x, y image coordinates: an updated count as a usage weight; a normal vector; and an accumulated distribution along the normal used to model the boundary points.

These identified trajectories are usually collected together into a database of routes by adding one trajectory at a time and using some measure of similarity, such as greatest overlap (see Fernyhough et al. 1996, 2000) or least maximum separation distance (see Makris and Ellis). Alternatively, given all the trajectories at once, Ng and Gong (2002a) create a similarity matrix by matching each trajectory against all the others, using dynamic time warping and the Levenshtein distance to obtain a measure of similarity, and then use the normalised cut to group the trajectories into self similar clusters, from which typical routes could be formed.

Map learning might also be applied continuously to capture dynamic contextual information, such as the effect of changing light and weather conditions illustrated in Stauffer and Grimson (2000), and the actions of scene participants. For example, in the road-traffic

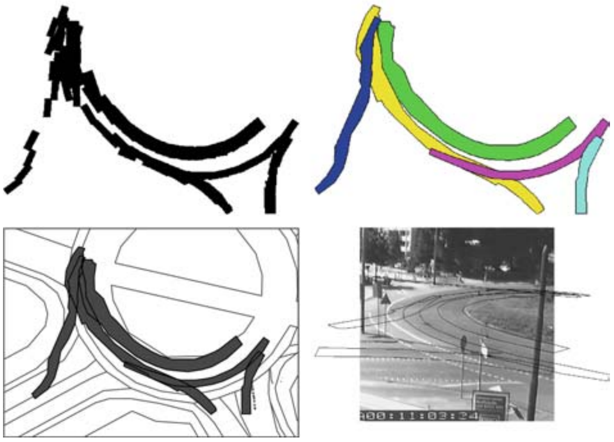


Figure 4. Learning paths. (a) Eleven different object trajectories that have been overlaid to illustrate the input data. The larger inner roundabout path is from the lorry shown in Figure 5. (b) The five learnt routes. (c) The routes in the context of the ground-plane. (d) The routes projected back onto the scene image.

domain a road user may have an accident which blocks a lane, or road works might alter the traffic flow. In the office domain, consider someone placing a large box in the middle of the typical path through the room. These actions break normal behaviour, and the model of typicality may need to be changed so that, for some temporal interval, it can represent the altered traffic flow due to these obstacles. Once the abnormal situation is resolved we may have to re-learn the old state again, and until this is done strange predictions may be produced such as going around an area previously occupied by a broken down car – an example of this is given by Sutton's (1990) blocking and shortcut problem.

As Regazzoni et al. (2001, p. 1356) explain, for surveillance systems to operate, in such highly variable scene conditions, robust methods that can automatically learn and adapt make "map learning" of normal activity patterns an emerging issue. And one that can provide a mechanism for distinguishing between what is normal and anomalous behaviour by flagging the least probable observed activity patterns.

2.3. *Robotics and motion planning*

Spatial representations in robotics (Schwartz and Yap 1987) and motion planning (Latombe 1991) mainly use geometric concepts (Kapur and Mundy 1989). The problems addressed tend to be concerned with control issues, such as finding paths in the presence of obstacles for a movable object or manipulator. However, the *surveillance problem* does not really make use of path planning, as we have no control over the activities of the observed actors who are dynamically planning their path through their environment. Computing the feasibility of each object continuing on its path dynamically seems impractical, both in terms of the observer having access to the intentions of each object and the dynamic nature of the environment.

An approach that is useful for full path planning, called "configuration space" (Lozano-Pérez 1983), involves "growing" the obstacles in the environment of a moving robot to reflect the extent of the robot's shape while the robot itself is simplified to point motion. However, in the *surveillance problem* we want to reason about interactions between objects with no single *central object of interest*. Treating all objects of interest equally is difficult to do using the configuration space approach, in part this is because each differently shaped object would require its own configuration space for each degree of rotation.

For example, see Latombe's (1991, pp. 373–384) description of multiple robots.

Two approaches related to path planning called "nonholonomic models" and "retracts" are worth noting. Nonholonomic models (Latombe 1991, pp. 403–451; Laumond 1993; also see Dorst 1998) provide a more realistic description of paths formed by objects, such as cars, by expressing the constraints upon possible velocity directions (differential motions). This provides an approach for generating better path predictions and checking the consistency of observed object motion.

Retracts (see Spanier 1966; Yap 1987; Canny 1988; Latombe 1991) provide a simplified model of space by using a function, $r : X \to A$ from a topological space X to a subspace $A \subset X$, where r is continuous and is the identity on A. When r exists A is called a "retract" of X, and the map r is called a "retraction" of X to A. Canny uses retraction to map 2D space onto a 1D subset, called the "roadmap", by using a simplified form of Voronoi diagram. For example, in the road-traffic domain, we could map the 2D space occupied by a road to a 1D line and describe the progress of the traffic on the road by points placed on this line. This would be useful for reasoning that only needed a simplified graph-like representation of traffic on the graph's 1D lines, however, in the *surveillance problem* we need to reason about the local spatial arrangements of the objects.

Unlike the work on robotics described here the *surveillance problem* is not concerned with controlling the objects it is observing. Thus, although the representations discussed here provide possible approaches to predicting what the observed objects might do next, and avenues to explore such as semi-algebraic sets,[15] the computational costs involved make most of these approaches unattractive.

2.4. *Graphics and solid modelling*

The primary objective of the surveillance system is not to display a graphical representation of the world, so most issues associated with computer graphics are not applicable to the surveillance task itself. However, there is a common concern with how to represent and reason about geometric information. For example, Mehlhorn (1984) and Preparata and Shamos (1985) describe geometric algorithms, Hoffmann (1989) describes issues related to constructive solid geometry,[16] and Kapur and Mundy (1989) provide links to work on vision and robotics. While these geometric approaches do not themselves help solve the *surveillance problem*, they can be useful as techniques for implementing a chosen spatial representation and for displaying 3D results. For example, Collins et al. (2001) describe how they have used a 3D computer graphic visualisation program, designed to support military interactive simulations, to model the CMU campus, including sensor placement, and then have inserted computer generated human and vehicle avatars that represent tracked and classified objects, to give real-time depiction of their results from any view point.

2.5. *Qualitative reasoning about motion and arrangements*

Weld and de Kleer (1990), in a useful section on reasoning about shape and space, cover various approaches on simulating a scenario for a given situation to find out what the outcome might be. The basic technique used typically employs "envisionment", which usually involves generating all the qualitatively distinct behaviours of a system for each possible initial state.

De Kleer's NEWTON program (1977) is an example of a topological spatial representation that uses envisionment. In NEWTON a rollercoaster ride is represented by an ordered sequence of regions such that, in each region, the values of "sign of curvature" and "tangent direction" are continuous; i.e., these two functions are used to define the extent of each region. The dynamic spatio-temporal aspect is described by "pre-compiling" all possible transitions between the regions into an envisionment which completely describes its world. Unfortunately, envisionment is not so good at handling unexpected cases or interrelations between multiple objects in complex domains. Thus envisionment would be useful for describing the individual paths of well behaved actors that repeat the same route, such as, trains and trams. However, in the everyday world, people often perform unexpected actions and it is precisely these actions that we want to identify. This makes envisionment, on its own, inappropriate for reasoning about everyday activities that exhibit some degree of freedom.

There seems to be some commonality between envisionment and the approach described by Fernyhough et al. (1996) – see Section 2.2. Perhaps these two techniques could be combined to identify and then incorporate regions that describe the likely behaviour at given places in the scene. The result from the observation of each vehicle/actor then, would not just be the learned routes but also some kind of generalised envisionment that represents where the various observed behaviours take place.

Envisionment has been used to model complicated domains, for example, the 2D possible activity of a bouncing ball (Forbus 1983; Forbus et al. 1987) and the CLOCK project (Forbus et al. 1991). The CLOCK project is more concerned with kinematic analysis, other examples are given by Gelsey (1987) and Kramer (1990, 1992). This form of spatial reasoning has requirements quite different to those of the *surveillance problem*, where the observed objects are not part of a closed kinematic chain in a mechanism, making kinematic analysis inappropriate.

A more formal definition of envisionment is provided by the axiomatic theory of Randell and Cohn (1989, 1992) who use a first order predicate logic formalism. Their theory describes a qualitative ordering, held in a lattice structure, that connects topological changes in pairwise relationships to enable allowable transitions to be determined.[17] There appears to be a substantial amount of theorem proving involved in this approach, which does not make it practical for a "real-time" problem like surveillance. For example, Cui et al. (1992) gives details of a simulator that implements this approach. Randell and Cohn base their formalism on the domain of regions (alternatively known as bodies or individuals) and use the connection between two regions as a primitive logical operator. This study of the logical properties of the relation of part and whole is called mereology. Other work on this includes: Taski's (1956) description of an approach using spheres where solids are constructed from spheres, like the layers of an onion; Clarke's (1981, 1985) calculus of individuals; and Asher's and Vieu's (1995) axiomatisation of mereotopology.

Freksa (1992) discusses various approaches to qualitative spatial reasoning and introduces the notion of qualitative orientation using an iconic representation that can be composed. Like the other approaches to qualitative reasoning described here this can help address those aspects of the *surveillance problem* concerned with dynamic scene object interactions. While these approaches are not so good at providing a model of the environment, the idea of abstracting out key features of the problem domain can be very useful. For example, on the VIEWS project, Toal and Buxton (1992) use a coarse grid representation to model the path swept out by the front of each vehicle and use this to detect following behaviour. This modelling of each vehicle's motion pattern can be thought of as an early version of the motion history image, described by Bobick and Davis (2001), that has been used, in conjunction with a coarse grid, by Ng and Gong (2002b) to learn where events happen in fixed camera video sequences.

2.6. *Linguistic and cognitive approaches*

At a higher-level of reasoning we can consider how language is used to describe geographic space (Talmy 1983; Herskovits 1986).[18] In spatial representation and reasoning linguistic concerns are sometimes seen as more important than the geometric and topological relations that they are trying to describe. It is easy to ignore the fact that spatial representation and reasoning can be performed without language (for example, Schöne (1984) and Gallistel (1990) describe how animals represent space) and that most of the linguistic generalisations are to allow communication by using common sense and commonly understood terms of reference that are made clear by their context. Recent work on spatial expressions (see Olivier and Gapp, 1998) illustrates some of the varied interrelationships between the expression of spatial information in both linguistic and non-linguistic forms. For example, Mukerjee (1998) presents a valuable survey of neat and scruffy techniques that have been used to represent space.

Additionally, work like Miller and Johnson-Laird (1976) and Lang et al. (1991) provides a linguistic approach useful for describing geographic space, communicating results and developing the terms from which conceptual descriptions can be formed. These ideas are developed further by Wachsmuth et al. (2000), who describe a system that integrates vision and speech understanding by using Bayesian networks to model the scene at different levels of detail. These levels are selected by Wachsmuth et al. (2000) to cover the terms people use to describe the objects in their application domain and the relationships used to distinguish each object from other scene features. This could provide a more natural interface to a surveillance system than that described by Collins et al. (2001) but, as shown by Chapman (1991) and Wachsmuth et al., the inclusion of this kind of instruction use needs to be an integral part of the system and not just a high-level afterthought.

Another example is the way that language is used to consider and communicate perceptions, giving an alternative perspective on how the perceived elements, entities and background features are arranged, conceptually, into correlated parts; Tversky and Lee (1998) and Talmy (1983) call this "schematization". Tversky and Lee explain how language influences perception by calling attention to certain features and relationships in a scene while disregarding others, and because of this the model of space used by language is more extremely schematized than that used by perception. However, having such a model could provide an alternative way to convey how people pass through a scene: rather than list geometric way points (e.g., GPS coordinates) of the route taken, it could instead be given in terms of the environmental background's reference objects and reference frames – see Tversky and Lee for more on how verbal and graphical route notation is used to communicate meaning. And of related interest is work that provides an initial, automatic annotation of object relationships in a scene.[19]

These examples illustrate the impact language has on how scene elements are denoted and referenced. Although these linguistic approaches are not so useful for representing the spatial data itself they could be used to augment a geometric model of the background environment and to describe the activities of the scene's foreground participants.[20]

Also of interest is Piaget and Inhelder's (1956) study on how children learn spatial concepts, using stages progressing from topology to Euclidean properties, corresponding to increasing age. This provides cognitive support for the use of topological reasoning in problem solving, which precede Euclidean relations in their simplicity. We will investigate the use of topological representations to determine if they offer any advantages in Section 3.

2.7. *Spatial decomposition*

Spatial data can often most easily be expressed in terms of a hierarchy based on enclosure. Rooms in a house, houses in a street, streets in a town, etc.. This has an obvious application in Geographic Information Systems (Laurini and Thompson 1992), and has also been used by Davis (1986) as part of his map learning program MERCATOR. Thibadeau (1986) describes an example room decomposition used by Heider and Simmel (1944) – these are compared in Howarth (1995). This decomposition reflects the faces of the room rather than the typical paths used by the participants. A similar approach is taken by the SOCCER project (see Blocher and Schirra 1995) reflecting the unconstrained movement of the players on the field/pitch. The different subdivisions of the pitch (e.g., field, right-half-field, right-penalty-area, right-goal-area) seem to be ordered by enclosure (for details, see Blocher and Stopp 1998).

Spatial decomposition also plays a prominent role in robot motion planning (see Section 2.3) as part of the formation of paths/graphs like the Voronoi roadmap. This process typically includes "cell decomposition" which is used to divide the available free-space into a finite number of coherent regions. Though as Thrun (1998, p. 60) points out, since the emphasis is on complexity analysis the resultant decompositions are sometimes odd-looking (for examples, see Latombe 1991). In contrast Thrun's map learning approach for indoor office/corridor environments identifies points on the Voronoi roadmap of locally narrow locations (such as doorways). These locations are then used to place region boundaries providing a decomposition that is more in tune with its environment.

Spatial decomposition is a useful approach that is independent of the spatial primitives used, allowing semantic properties to be attached to spatial elements (such as rooms, houses, etc. in the example above). And, although spatial decomposition only addresses part of the representation problem, it can make an important contribution to how space is modelled by providing an overall structure that corresponds to task relevant environmental features.

2.8. *Intermediate-level vision*

The representations of space used in vision research are not often concerned with modelling geographic or natural features, however, there are exceptions such as Pentland's (1986) use of superquadrics[21] to describe natural forms, the scene reconstruction work described in Section 2.1, and the description of human motion by Badler and Smoliar (1979), O'Rourke and Badler (1980) with related work described by Aggarwal and Cai (1999), Gavrila (1999). The majority of representations are concerned with supporting low-level vision – for example, see Marr (1982) and Horn (1986). These approaches use geometric and topological spatial representations to describe the image features in the image-plane. One notable example is provided by Fleck (1988a, b, 1991) with her topological approach to representing digitised spaces for both edge detection and stereo matching. In general there is no commonality between the use of the spatial representation in low-level vision and higher-level vision. As Ullman (1996) explains: Low-level vision is concerned with preliminary processing of the image and is spatially uniform and parallel where as higher-level visual processes are applied to information from a selected portion of the image with the objective of interpreting what is seen. However, as we describe later, there can be commonality in the representation used.

Intermediate-level vision typically involves object recognition such as that shown in Figure 5, where a simplified 3D model of the known static environment which includes road outlines and other important features (such as the posts shown in Figure 5(b) that occlude parts of the moving scene objects) enable the various 3D models to be fitted to the scene objects facilitating their recognition. And these 3D models may use the solid modelling techniques described in Section 2.4. For further details see Koller et al. (1993), Murray and Buxton (1990), Nagel (1988), Tan et al. (1998), and Worrall et al. (1991).

One of the most wide ranging models of space is given by Fleck (1988a, 1996) which, in addition to describing how to represent digitised spaces for both edge detection and stereo matching, also contains applications of her topological approach to natural language semantics and qualitative physics making it a natural candidate for extension to the *surveillance problem*.

2.9. *Analogical representation*

In general, all spatial representations that model reality using pictorial or diagrammatic means are analogical. Indeed, it seems likely that, this analogical component needs to be present in order to make them

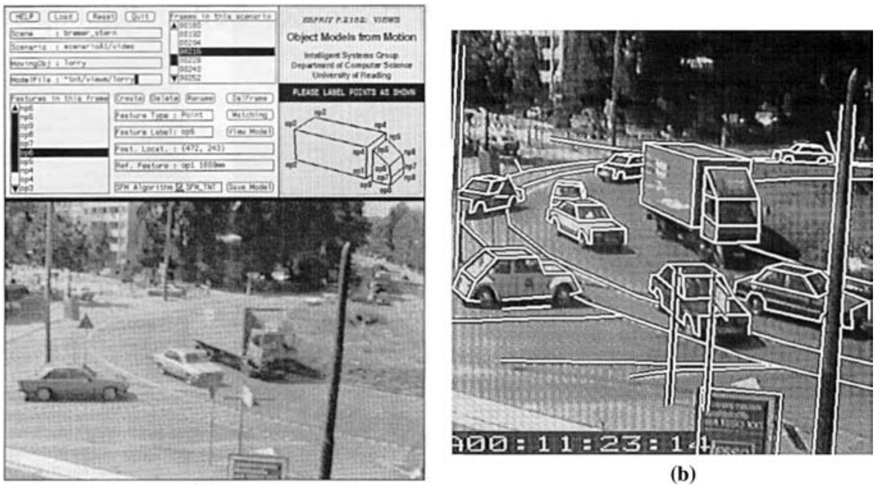

Scene : bremer_stern
Scenario : scenarioA1/video
MovingObj : lorry
ModelFile : ~tnt/views/lorry
Frames in this scenario
00180
00192
00204
00216
00228
00240
00252
ESPRIT P.2152: VIEWS
Object Models from Motion
Intelligent Systems Group
Department of Computer Science
University of Reading
PLEASE LABEL POINTS AS SHOWN
Features in this frame
Feature Type : Point
Feature Label: op6
Feat. Locat. : (472, 243)
Ref. Feature : op1 1000mm
SFM Algorithm ☑ SFM_TNT

Figure 5. Interactive object modelling. Picture (a) shows the work window, and (b) shows fitted models. These illustrate the work done on the VIEWS project by the Intelligent Systems Group, Department of Computer Science, University of Reading. From Corrall and Hill (1992, p. 27). Reprinted from GEC-Review with permission.

generally understandable. Sloman (1975, 1985) gives a good explanation of what can be an analogical representation, describing how, in an analogical/modal representation, the constituents and their relations must have some sort of correspondence with the represented reality (or the thing the representation is being analogical too), and how this differs from a Fregean/symbolic/amodal representation where there need be no such correspondence. This analogical correspondence is most obviously present in visual perception – see, for example, Kosslyn (1994, p. 404), Barsalou (1999), Johnson-Laird and Wason (1977) – and is pertinent to our *surveillance problem* because we are representing spatial data to be used in a visual task. This gives us two different levels of analogical correspondence to reality. At one end is the spatio-analogical model of the environment, e.g., such as a map, diagram or picture; see for example Monmonier (1991), Barkowsky and Freksa (1997). And at the other is the cognitive representation and processing used to perceive the environment – which is, as Pylyshyn (1984, pp. 193–223) describes, a much more difficult problem.[22]

The spatial representation being developed here is not necessarily a model of the internal cognitive representation people use, neither is it an external representation, such as a map or diagram, to help people reason about some cognitive task. Instead we are developing a knowledge representation that is to be used by another computer program. Making the *surveillance problem* more an engineering task, where we are looking for a parsimonious computational solution.

This kind of computational spatial representation is present in analogical models of physical systems such as Forbus et al.'s (1991) CLOCK project (see Section 2.5) and Gardina and Meltzera's (1989) 2D pixel array, naive physics simulation of strings and liquids, and Yip and Zhao's (1996) imagistic problem solvers; and also the learning of vehicle trajectories described in Section 2.2.

Other work that takes account of cognitive issues as well includes Funt's (1980, 1983) WHISPER program which manipulates, in its "mind's eye", a direct analogue of a physical blocks-world-like situation to detect potential instabilities and collisions; and Steels' (1988, 1990) "internal representation that is similar and close to sensor outputs", made from a combination of cellular automata (see Toffoli and Margolus (1987) – basically the same operation is applied in parallel to the content of all the cells in a grid) and the biochemical metaphor of reaction–diffusion, and which Steels uses to do things like search for objects in a map and avoid obstacles; and Glasgow and Papadias

(1992), and Glasgow's (1993) computational model of imagery that uses three different representations of spatial data (long-term-memory using frames (Minsky, 1975); visual image data using 2D or 3D occupancy arrays; and a qualitative model of topological relationships held in a symbolic array to both structure space and give a metric ordering) that all have some kind of equivalence relationship with each other, with what they are modelling, and with Kosslyn's theory of mental imagery.

These illustrate the diversity of analogical representations and also their common feature where each provides a selective correspondence to reality, capturing those key spatial properties relevant to the particular problem being solved.

2.10. *Spatial uncertainty*

A spatial model for a visual surveillance system may also need to represent the increase in noise due to distance from the camera. Two approaches appear relevant: (1) the work of Durrant-Whyte (1988), who describes geometries designed for reasoning with uncertain data; and (2) "tolerance space" which could be adapted to provide a graduated tolerance linked to the perceptual acuity of the scene. Hayes (1985) describes how indistinguishability is a tolerance relation and how a space defined by such a metric is called a tolerance space. Tolerance provides a natural notion of distance between "qualities", describing the smallest number of "steps" by which one quality can be transformed into another, with each step being invisible under the tolerance. For example, consider the colours Red and Blue, and how we could transform from one to the other in such small steps that the difference between the steps is not distinguishable. The tolerance relation was initially developed by Poincaré (1958), who used it to provide a way of expressing "if A is near B and B is near C, then A is not necessarily near C". For example, using colours again, consider A representing Red, C representing Blue, and B representing Purple.

The ideal of displacing Euclidean geometry in the physically small has been investigated by Kaufman (1991), Davis (1988, 1989), Poston (1971), Dodson (1974), Roberts (1973), Zeeman (1962). For example, Fleck's topological representation uses something similar called "error neighbourhoods" when comparing the values at two cells – see (Fleck 1988a, p. 51; 1988b, p. 77). The two values are indistinguishable if their error neighbourhoods overlap. The tolerance space approach could be adapted to represent a gradient of indistinguishability that models the effect of depth from the camera. As shown in Figure 1 this is not depicted in the ground-plane projection. By adding tolerance neighbourhoods to both static ground-plane features and dynamic object data in accordance to this depth gradient, points close to the camera would be clearly defined while those much further away would only have a coarse, fuzzy description.

Rather than describe the whole space a related approach would be to just apply the tolerance space to a single object of interest in order to determine its relationship to other objects. By fitting a conceptual "field" to our reference object we can establish the relative proximity and orientation of any other relevant object, generating a continuum measure for a qualitative state, such as, infront. Mukerjee et al. (1998, 2000) give more details about continuum measures, which have also been used in the SOCCER project (see Schirra and Stopp 1993) to generate spatial expressions. This field representation is similar to Schöne's (1984, p. 31) model of the dissipation of scent and the potential field representation described by Latombe (1991, p. 19). This approach has also been used in HIVIS-WATCHER to provide a proximity measure around each object's posebox, i.e., the model-matcher results illustrated in Figure 5(b), as described in Howarth (1998, p. 25).

2.11. *Discussion*

We have considered spatial representations both in terms of how they are viewed in the everyday world and how they have been viewed in a broad range of artificial intelligence applications. These outline the scope and various purposes to which spatial representation is put, showing how "purpose" is important in computational systems both natural and artificial, with excess functionality typically present to tailor the representation to some particular spatial reasoning objective. Often this concerns events and attention – the two closely related parts of the *surveillance problem* introduced in Section 1, but not directly addressed in this paper except where they overlap or interrelate with the issues discussed here – for example, Table 1 has columns for the perceived events of the participants and the attentional, task-level behaviour of the observer.

This survey and Section 1.4 have shown some of the ways environmental context is useful for interpreting observed behaviour. For example: large scale space can provide details about the wider location; map learning can provide typical routes that describe previous use of

Table 1. Suggested relevance.

Spatial representation schemes discussed: section titles	Relevance to *surveillance problem* issues								
	Off-line	On-line	Single camera	Multi camera	Simple detection	Events	Observer	Interactive multi media	Site modelling
Large scale space	●			●					●
Map learning	●		●	●		●	●		
Robotics and motion planning	●					●			●
Graphics and solid modelling	●								●
Qualitative reasoning about motion and arrangements	●	●	●	●		●	●		
Linguistic and cognitive approches		●				●	●	●	
Spatial decomposition	●		●	●			●		
Intermediate vision		●	●	●			●		
Analogical representation	●	●	●	●		●	●	●	
Spatial uncertainty	●	●	●	●		●	●		

the environment; linguistics and cognitive approaches provide a labelling or interpretation of reference objects and reference frames; and scene decomposition adds another form of structure and organisation. These all describe ways that knowledge about the structure of an environment, and what can take place within it, can help solve the *surveillance problem*.

Table 1 gives some of the possible groupings of topic areas looked at in this section. Not all of them are necessary for every surveillance application. For example, a simple detection system might not need any details about the environment it operates in. It is for applications which require an interpretation of activity, situated in its environment, that such context becomes relevant. This can be separated into data only available at run-time and data that can be acquired before any surveillance task is performed; we call these "on-line" and "off-line". And while there is likely to be overlap between these two, the off-line topics are those that are useful for constructing a spatial database or to enhance a site-model.

Obtaining the necessary context appears to be a simple matter of having all the background knowledge that makes interpreting each trajectory and activity meaningful. It is surprisingly effective how placing results back in the context of the image makes their meaning become clear. However, this understanding is not provided by the computer program but by the reader. The ease of supplying context in this way (see Figure 4 for example) should not hide the potential benefits knowledge about environmental context can provide activity interpretation. Extracting this contextual knowledge can be difficult as it is often connected to general knowledge and common sense reasoning as shown by Draper et al. (1996), Pentland (1986), and Strat and Fischler (1991) in their discussions on how to recognise and describe natural forms like trees, and man-made objects like buildings. Other relevant techniques, like labelling image features based on texture and colour and image position – see Adams and Williams (2002), Feng et al. (2002), and Duygulu et al. (2002) – which learn a mapping between image features and their names, provide another link to the linguistic approaches of Section 2.6.

This survey has covered a number of areas, each of which has a different relevance to the *surveillance problem* and some more so than others. Perhaps an ideal activity interpretation system might include aspects from all of these, but it is more likely that a subset of these elements would be selected to fulfil some particular task.

For example, the multi-camera installation of Collins et al. (2001) uses the GPS world wide coordinate system as a framework for their large scale site model and computer graphics to provide a real-time visualisation.

Howarth and Buxton (1992) use spatial decomposition, analogical representation, ideas from qualitative spatial representation and Fleck's model of space (see Section 2.8) to describe the Bremer Stern roundabout (see Figure 1).

Toal and Buxton (1992) combine a number of reviewed areas in their VIEWS project work: occlusion reasoning, camera field of view, coarse analogical grid map; and by depicting the camera's field of view on the grid map they are able to detect occlusions and from this construct a continuous path for each occluded object.

Even though Kuipers' (2000) spatial semantic hierarchy was developed for robot exploration of large scale space and map-building and uses ideas from human cognitive map research, it presents an important example of how multiple interacting, qualitative and quantitative, topological and metrical, spatial representations can be integrated in a common framework that can express states of partial knowledge and deal robustly with uncertainty.

This idea of using a semantic hierarchy is also present in Forbus et al.'s (1987) description of FROB that separates the spatial representation into conceptual layers. They use different layers for the solid form of the environment (walls of a 2D open top box); a symbolic description of points, lines, and regions (inside the box, above and either side) for the metric diagram; and a history level describing where the ball has bounced. Drawing from the work described in Section 2, we could have a conceptual layer for each of: the image-plane; the ground-plane outline or 3D visualisation; a labelling of reference objects and reference frames (from linguistic and cognitive approaches); the learnt routes, so we can say how a subsequent trajectory relates to those we have seen (from map learning); each object's posebox outline and a measure of mutual proximity (from spatial uncertainty); a coarse grid for simplified motion history, etc. (from qualitative reasoning). Plus another layer for any of the contextual representations described in Section 1.4. Each of these layers give a different aspect about the spatial occupancy of the moving objects in the environment. Also, this decomposition into a framework of functionally distinct conceptual layers provides an enhanced representation of the real-world that is more than just a 3D reconstruction of the site, and being an analogical, mental model view of what the observer "knows" about the environment it can integrate perceived reality with internal cognitive representations. Some of these notions were used on the VIEWS project.[23]

Common to all approaches has been the use of a spatial representation although no single set of spatial primitives is evident. We take note of the priority that Piaget and Inhelder (1956) place on topological reasoning, and will take a closer look at Fleck's (1988a) model of space because it supports both topological and Euclidean reasoning. In Section 3 we describe a range of spatial representations that can support the requirements of surveillance introduced in Section 1.5.

3. The Cellular Spatial Model

As illustrated in Figure 1, in the *surveillance problem*, the representation of space has been simplified to provide a 2D model of the ground-plane and scene objects. Although reference is sometimes made to 3D forms they are not fully addressed and are outside the scope of this paper. The *surveillance problem* includes both metric and topological data from which we want to determine things like: the distances and angles between objects, and the spatial location that each object occupies. Consideration of how the surveillance system will be used impacts upon the design of the spatial model. Also, the implementation needs to address issues of efficiency by eliminating unnecessarily repeated calculations. These issues are particularly apparent with boundary representations (e.g., Davis, 1986; Hoffmann, 1989) where topological relationships need to be determined a new for each test via explicit calculation. It makes more sense to use a representation that more directly describes the topological information.

3.1. *Space filling cells*

Here we investigate the work of Whitehead (1949) and Fleck (1988a, 1996) and describe how we can use topological relations (like closure, interior, boundary, separated and connected) in conjunction with Euclidean distance. The basic element is called the **n-cell** where the n denotes the dimension of the cell. A 0-cell is a point or vertex, a 1-cell is a line, an edge or face of a higher dimensional object. Some example 2-cells are shown in Figure 6(a). This cellular representation provides an abstract model that can be implemented in a variety of ways and can underly most of the spatial representations outlined above. Informally,

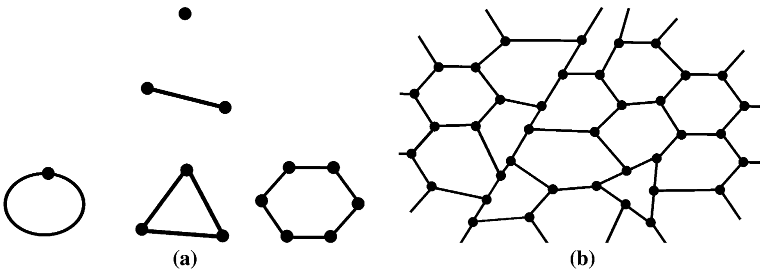


Figure 6. Example cells. (a) From top to bottom, these are a 0-cell, a 1-cell (with two 0-cells) and three 2-cells (going from left to right: the first has one 1-cell and one 0-cell, the second has three 1-cells and three 0-cells, and the third has six 1-cells and six 0-cells). (b) A regular cell complex showing that a regular pattern is not necessary in a complex of regular cells. Here the 2-cells all have six 1-cells and six 0-cells.

a cell is a blob of space, having some fiducial size. The cells do not need to have the same size but they will all be of some limited range of sizes according to the granularity of space used. The basic idea is that we will build our spatial representation out of these cells.

Figure 6(b) shows how cells can structure space, providing a framework to which we can attach information. In Section 3.1.2, we introduce the term **regular cell complex** which outlines Whitehead's (1949) central idea of cellular topology that enables us to describe a collection of disjoint cells such as those shown in this figure.[24] First we begin by placing cellular topology in a historical context, for example, Lefschetz (1970) describes how Poincaré introduced the concept of complexes and the highly elastic algebra that goes with it.

3.1.1. *Simplicial*

Maunder (1970, p. 61) says that the study of simplicial complexes is one of the oldest parts of topology and dates back at least to Euler. The boundary of an n-simplex consists of the union of the $(n-1)$-simplices, called faces. For example, the boundary of a 2-simplex $<a, b, c>$ is made up of the three lines $<a, b>$, $<b, c>$ and $<c, a>$. In simplicial topology (see Alexandroff 1961) the boundary operator works on a vector space (over integers or some other algebraic field) generated by using the "$< \cdot >$" symbol for each simplex and using orientation reversal as negation. The boundary of a 0-simplex is the zero-vector having only one orientation; i.e., $\delta(<a>) = 0$. The boundary of a 1-simplex is the vector difference of its two 0-simplices $\delta(<a, b>) = <b> - <a>$. The boundary of a 2-simplex is the closed sum of its oriented edges $\delta(<a, b, c>) = <a, b> + <b, c> + <c, a>$. Alexandroff (1961, p. 20)

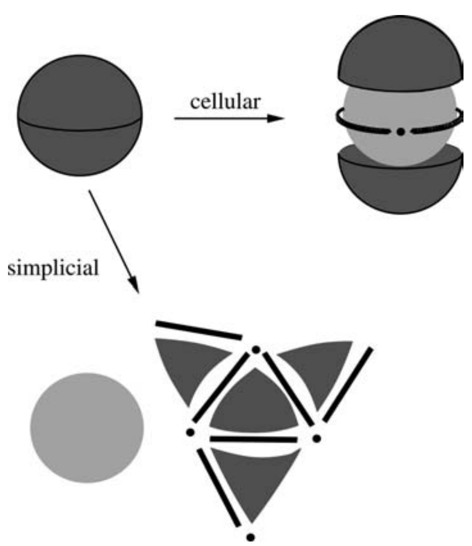


Figure 7. Representing a sphere in simplicial and cellular topology. For a cellular realisation a solid sphere needs at least one 3-cell, two 2-cells (the northern and southern hemispheres), one 1-cell (the equator) and one 0-cell (where the ends of the equator meet), which gives six cells in total. For a simplicial realisation it would require at least one 3-simplex, four 2-simplices, six 1-simplices and four 0-simplices, which is fifteen in total, to describe the 3D oriented simplex of a tetrahedron, providing a topological equivalent to our sphere.

explains how a simplicial complex is composed of these oriented simplexes $<a_0, \ldots, a_n>$.

3.1.2. *Cellular*

Herring (1991, p. 334) explains that the big difference between simplices and cells is for visualisation and storage. For example, as shown in Figure 7, a cellular realisation can be made with fewer cells than a simplicial one. To briefly explain how to define the structure of a particular space, by using regular cell complexes we first need to stipulate the separation axiom of our topological space by using definitions of neighbourhood and of Hausdorff space[25] – see, for example, Bredon (1993, pp. 4–13), Maunder (1970, p. 15), Munkres (1975, pp. 94–98).

- If x is a point of X, each **neighbourhood** of x is an open set that contains x.
- A space X is **Hausdorff** if for each pair of distinct points $x_1, x_2 \in X$, they have neighbourhoods that are disjoint, i.e., there exist neighbourhoods U_1, U_2 of x_1, x_2 respectively, such that $U_1 \cap U_2 = \emptyset$.

In the usual topology on the real line $\Re$, a subset is said to be **closed** if its boundary points are included (e.g., $\{x \mid 3 \leq x \leq 5\}$) and **open** if the boundary points are excluded (e.g., $\{x \mid 3 < x < 5\}$). To explain what a boundary is, let A be a subset of a topological space X, the **interior** of A (denoted $\overset{\circ}{A}$) is the union of all open sets contained in A, and the **closure** of A (denoted $\overline{A}$) is the intersection of all closed sets containing A. Then we can say that the **boundary** of A is $\overline{A} \cap \overline{(X - A)}$, i.e., the intersection of the closure of A and the closure of the complement of A. The boundary and interior of A are disjoint and when unioned together form the closure of A.

Fritsch and Piccinini (1990, p. 1) explain that the standard models used in cellular topology to describe cells are balls and balloons. B^n is the unit **ball** of dimension n. The boundary of B^n is the **sphere**, is of dimension $n - 1$, and is like a balloon or skin of a ball. A **cell** of dimension n is homeomorphic with B^n, and an **open cell** of dimension n is homeomorphic with the interior of B^n. By homeomorphic we mean that they are topologically equivalent.

Now we are in a position to define what a regular cell complex is. As described by Whitehead (1949, p. 95), Munkres (1984, pp. 214–221) and Massey (1980, pp. 76–104): a **regular cell complex** or **CW-complex** is a space X and a collection of disjoint open cells e_α, whose union is X such that:

1. X is Hausdorff.
2. (**Closure-finiteness**) For each n-cell e_α of the collection, there exists a continuous map $f_\alpha : B^n \rightarrow X$ that maps the interior of B^n homeomorphically onto e_α and carries the boundary of B^n into a finite union of open cells, each of dimension less than n.
3. (**Weak topology**) A set A is closed in X if $A \cap \overline{e}_\alpha$ is closed in $\overline{e}_\alpha$ for each α. Which means that each cell, $\overline{e}_\alpha$, of X is contained in a finite subcomplex of X.

Herring (1991, p. 332) points out that the use of "complex" here has nothing to do with complex numbers, instead it is to do with how a whole is made up of interconnected parts such that an n-dimensional complex is a collection of n-dimensional elements (e.g., n-cells) sharing common boundaries. For example, as Munkres (1984, p. 217) explains: a **subcomplex** L is a subspace of some CW-complex X that equals a union of open cells of X, such that for each of these open cells its closure is also contained in L. L is closed in X and is itself a CW-complex. Also an n-**skeleton** $X^{(n)}$ of X is a subcomplex of X such that it contains a union of open cells that have at most dimension n.

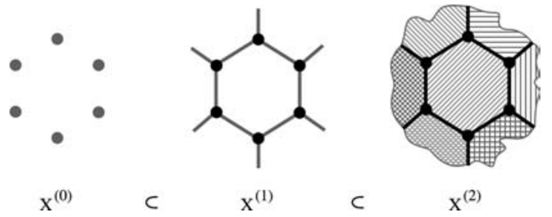


Figure 8. Skeleton.

Intuitively, a regular cell complex is a space which can be considered as a union of disjoint "open cells" providing a tiling of space with $X^{(0)}$ being a discrete set of points, the 0-cells. And, if $X^{(n-1)}$ has been defined then $X^{(n)} = X^{(n-1)} \cup_f Y$ where Y is the disjoint union of copies of B^n, one for each α, and $\cup_f$ is the collection of maps f_α as described above and in Figure 8. Further details are given by Fritsch and Piccinini (1990). We have now covered the material necessary to describe Fleck's extension to cellular topology.

3.1.3. *Adjacency sets*

Fleck (1988a, b) shows that a structure-preserving mapping between the adjacency structures of two cell complexes provides a homeomorphism between their underlying spaces. This approach is simpler than building homeomorphisms directly and has been used to fuse stereo image pairs, while the structure itself can provide a nice way of representing spatial models. Adjacency sets are used to specify the adjacency relationships between cells and, as described in

Lemma 5 (Fleck, 1988a, p. 389), an **adjacency set** must comply with the following:

1. Every $(n-1)$-cell must be a face of at least two n-cells.
2. There is a fixed dimension n, such that each cell is either an n-cell itself or it is a lower dimensional face of an n-cell.
3. The intersection of any set of cells must be exactly one cell or empty.

The first clause ensures that neighbouring regions have a common edge and that each vertex has at least two edges.[26] The second clause ensures that space is of a constant maximal dimension. The third clause ensures that two neighbouring regions can *only* share *one* edge. Fleck (1988b, pp. 76–77) gives an illustrative example and describes how the information contained in an adjacency set structure can be used to build lower dimensional skeletons, like those shown in Figure 8, and attach higher-dimensional cells to them.

3.1.4. *Inter region boundaries*

It is not always easy to describe the boundary between two objects. Often we have the question of which object *owns* the boundary? For example consider the boundary between a table and a mug placed on it, does the boundary belong to the mug, the table, both, or neither? Sometimes the boundary is arbitrarily assigned to one or other of the objects or, in the case where one of the objects is the background, boundary ownership might be treated differently. Alternatively we might have the boundaries overlap by one point so that both objects own the boundary. None of these approaches provide a good representation of the boundary between two objects. To solve this problem Fleck (1988a, p. 394) extends the definition of regular cell complex to include boundaries: A regular cell complex **with boundaries** is a regular cell complex together with a list of cells called the **boundaries** of the complex, so that the space we are modelling is represented by both a set of space filling cells that structures the space, and a list of cells endowed with the boundaries.

When reasoning about cell complexes we are primarily interested in this definition of boundary features. We can use either an open-edge or a closed-edge boundary model throughout our spatial representation. In the open-edge model of boundaries, points corresponding to boundary adjacency sets are simply deleted from space. This means that the cells to either side of the boundary are now next to each other but are no longer connected. The closed-edge boundary model is similar but here points are added to "close" the edges, making them look like closed subsets of real space. The new points on either side of the boundary are right next to each other but distinct.

3.1.5. *The new open model*

Fleck (1996, p. 10) describes an open model requiring that:

- The boundaries are a (topologically) closed set of points.
- The boundaries in any bounded subset of $\Re^n$ are isotopic to the union of a finite number of semi-algebraic sets.

As described above we delete all points in the boundaries from $\Re^n$ so that they no longer form part of any region. Semi-algebraic sets were introduced in Section 2.3 and are discussed further in Section 3.3.2 as being a possible option for representing cells. Isotopy provides a stronger constraint than homeomorphism such that given two spaces, one space can be continuously deformed into the other – see Fleck (1996, p. 4) and Rouke and Sanderson (1982). This provides an alternative to the adjacency set approach.

3.1.6. *Summary*

This has introduced the cell as a spatial primitive and sketched its history in topology. Section 2.8 described how Fleck has applied the cellular topology to a wide range of applications which illustrates how this representation is more general purpose than most other spatial representations. Also, because topological concepts are present in most of the computational approaches we looked at the cellular topology can be thought of as a common representation.

Our next objective is to consider how this cellular representation can be implemented.

3.2. *Polyhedral tessellations*

The *surveillance problem* does not require a perfect model of the real world, so curves can be approximated by polyhedra reducing the implementations complexity. A polyhedron can be composed from a set of 1-cells, the connecting 0-cells and the 2-cells that the boundary 1-cell faces enclose. We can describe this as:

> A n sided (simple) **polyhedron**, $\mathcal{P}$, is composed from a set of n 1-cell faces, $\mathcal{L}_i$, such that $\mathcal{P}=\sum_{i=1}^{n}\mathcal{L}_i$, and that no members of the set $\{\mathcal{L}_1\ldots\mathcal{L}_n\}$ cross (intersect).

This definition is important because it provides a geometric framework within which to model the 2D ground-plane and also because it acts as a suitable interface to any space structured using a regular cell complex. The definition does not exclude polyhedra with concave vertices, although it is worth noting that such polyhedra tend to make topological tests for occupancy more complex. The implementation should not affect the geometric map specification given in terms of polyhedra. There are two approaches for translating from a specification in terms of polyhedra to one in terms of regular cell complex which are distinguished by how they structure the underlying metric space. These approaches are called “regular” and “irregular”.

3.2.1. *Regular*

A regular, uniform subdivision of space is composed from a single repeated regular shape rather like a regimented pile of sugar cubes.[27] We can apply the same ideas to tile the 2D plane, using one of the regular polygons: equilateral triangle, square (e.g., pixels in a “raster”) or hexagon (for a pixel level example see Bell et al. 1989); to give a regular tessellation of a pattern of congruent regular polygons, all of one kind, filling the whole plane.[28] The main problem with this approach is that the structure of space does not always match the object boundary that is being represented and results in "jaggyness". However, this approach to structuring space is very attractive because of its simplicity, directly describing the spatial extent of an area in terms of occupied cells, which gives this approach the name "occupancy grid". It also marries well with implementations in terms of bitmaps, 2D arrays or hash-tables of used addresses. When the granularity is too large, severe distortion and loss of detail can result. The solution is to use a finer granularity, however, this can cause a major storage problem. This storage problem can be greatly reduced by using quadtrees (Samet 1984) or tesseral arithmetic (Gargantini 1982; Bell et al. 1983; Coenen et al. 1998). Both techniques can provide savings on storage of silhouette like shapes, where representation of an area is important. All members of this family of occupancy grids support the efficient implementation of set operations.

3.2.2. *Irregular*

In the regular structures above the 2D polygons are the smallest elements. We could also use the 1D edges of each 2D polygon however, this would still exhibit the same representational problems outlined above, i.e., jaggyness. This problem can be addressed by removing the "regular structure constraint" to allow an irregular tessellation where the structure of space, formed by the tessellation, can be chosen to comply with the spatial form of the environmental properties – it could also be used to provide a regular tessellation so can be thought of as the superset of all tessellation methods. This means that edges in the environment are represented by cell faces in the structure used to describe space. The presence of a matching structure removes the problem of cell granularity, as long as all the required cells are present in the supporting structure of space. There are two situations in which tessellations arise:

1. Given a set of points find a tessellation.
2. Given a set of non-intersecting polyhedra find a finer tessellation.

An example of the first type is connecting the measured spot height elevations of a landscape. The spatial representations used in the *surveillance problem* are of the second kind where we use the known environmental structure (described in terms of polyhedra) to influence how the underlying cellular representation is to be tessellated.

3.3. *Tessellation techniques*

Many geometric algorithms perform better on convex point sets. Thus, it is expedient to decompose concave polyhedra into combinations of convex polyhedra. The most primitive is the triangle, for which there are a number of triangulation algorithms, some of which are discussed next.

3.3.1. *Triangulation*

Here we discuss why some of the standard approaches are not appropriate, point out some efficient algorithms, and identify interesting approaches to optimal decomposition both in terms of the number of polygons and the fiducial size of the cells produced.

In an irregular tessellation we may have a set of required edges between vertices that describe the extent of environmental features (e.g., walls and typical paths). This set of required edges makes some tessellation techniques such as Delaunay triangulation (Guibas and Stolfi 1985; Sloan 1987) inappropriate, because they do not allow special edges to be retained. Although the basic Delaunay triangulation is inappropriate, Sapidis and Perucchio (1991) describe an extended Delaunay triangulation algorithm that could be used, since it is able to triangulate the space inside a specified polyhedron.

An alternative to Delaunay triangulation is its dual, called variously a Voronoi diagram or Dirichlet or Thiessen tessellation. A Voronoi diagram produces a tessellation of convex polygons, and can be performed so that it retains the special edges. This tessellation is likely to generate a large number of small polygons that require further tessellation to produce a triangulation. The benefit of this approach is the generation of the medial axis for the special edges providing their retract – see Section 2.3 and Latombe (1991, pp. 169–174) and Fortune (1987) for further details. Both Hobby (1993) and Gold (1992) give examples of a Voronoi tessellation providing line segments that form a polygon set of special edges. The properties of the Voronoi diagram can be described in relation to the problem called *loci of proximity* because its solution results in a data structure that represents the nearest neighbour of each point – see Preparata and Shamos (1985, pp. 204–225). This can be described as follows: Given a set of n points in a plane called **sites**, for each point; $p_i \in$ sites; there is a locus of points (x, y) in the plane that are closer to p_i than to any of the other $n-1$ sites. (Gold (1990, 1992) describes how this can be pictured as the result of expanding a balloon around each site until all the balloons are balanced by their neighbours. Each locus is an individual balloon boundary.) This collection of loci structures space into polygonal cells (2-cells) with a 1-skeleton formed by the loci boundaries. The **Delaunay triangulation** results from joining two points; $p_i, p_j \in$ sites; if they share a boundary edge (i.e., a 1-cell in the 1-skeleton). The **Voronoi diagram** is the 1-skeleton provided by the boundary edges.

There are other triangulations that have different constraints and which are more appropriate for implementing cells, in part, because they provide faster linear time triangulation algorithms. For example, Kong et al. (1990) describe an algorithm that is linear for simple polygons that have few concave vertices – this is the one used in Howarth (1994). For more details see El Gindy and Toussaint (1989), who discuss how varieties of polyhedral shapes affect the triangulation problem. Various other algorithms are given by Preparata and Shamos (1985), and Mehlhorn (1984).

As an alternative to forming triangulations, we might only need to form a combination of convex polyhedra (which could then act as a preprocessing stage to linear time triangulation). There are two approaches we can take. The first is to add new points, called Steiner points, which are vertices inside a polygon P but not on the boundary of P, as described by Chazelle and Dobkin (1985). The alternative approach is *not* to add new points and algorithms for this are given by Green (1983), Tor and Middleditch (1984). Keil and Sack (1985) provide a more detailed comparison, but basically the addition of Steiner points allows an optimal decomposition to be performed.

Once we have completed a tessellation we might find that some of the cells are to large, one option is to use a Barycentric subdivision (Alexandroff, 1961), which operates on convex shapes to reduce the granularity of the component parts by triangulation.

These tessellation techniques provide an insight into the range of approaches for representing cells. Under the current requirements of the *surveillance problem* much of this functionality would be redundant. However the use of retracts, via the Voronoi diagram, is appealing and may prove useful during a knowledge acquisition phase as a framework for modelling a "roadmap" of typical behaviour.

3.3.2. *Cell decomposition*

Other relevant approaches include exact cell decomposition (see Section 2.7) where the structure of space is described using a set of algebraic equations.[29] For example, the open unit disc shown in

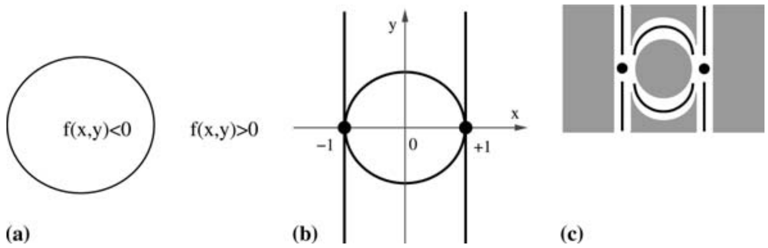


Figure 9. An example based on the unit disc. (a) The inside and outside of a cell, where $\mathbf{f}(x, y) = x^2 + y^2 - 1$. (b) A Cylindrical Algebraic Decomposition of the unit disc, adapted from Latombe (1991, p. 240). (c) An example "exploded" illustration of the cells formed by this decomposition.

Figure 9 can be described by the semi-algebraic set $S = \{(x, y) \mid \mathbf{f}(x, y) < 0\}$ where, as shown in Figure 9(a), $\mathbf{f}(x, y) = x^2 + y^2 - 1$, so that S includes the regular points in the interior of the disc. In addition the space outside a 2D shape can be partitioned into a finite collection of semi-algebraic cells which, as shown in Figure 9(b) using the same polynomial $x^2 + y^2 - 1 = 0$ for the unit disc, can be recursively decomposed into thirteen cells (two 0-cells, six 1-cells, five 2-cells) by using Collins' (1975) decomposition algorithm that employs a method based on Tarski's (1948) approach for deciding the satisfiability of Tarski sentences – also see Latombe (1991, pp. 225–242).

3.4. *Implementation options*

Once we have our cellular decomposition, we then have the problem of how to store the points, lines, and areas in various data structures. While Fleck's approach of separately storing the structural mesh and the list of boundary cells is a viable option. We can also form the spatial model from polygonal regions that represent these important edges and then make use of the available decomposition techniques to create a suitable structural mesh. Examples include: the bitmap arrays of Section 3.2.1; and the triangular cells described in Section 3.3.1. Alternatively, we could adopt the winged edge model of Paoluzzi et al. (1993) as used in their implementation of simplicial complexes. Another promising approach is described by Günther (1988, 1991), who uses the simplicial approach given in Section 3.1.1 to combine edges in the form of hyperplanes so that they model the polyhedra without the need for specifying vertices. Or we could do something similar but use semi-algebraic sets instead of hyperplanes (see Dumortier et al. 1997). Also, since we are talking about combinations of cells, we could treat the cells as constructive solid geometry (CSG) primitives (see Hoffmann 1989) so that a combination of cells would be represented by a binary tree with union operations at the nodes. These examples provide an insight into the range of options available.

In addition to our cellular model, the spatial data itself may express hierarchical properties that we would like duplicated in the spatial model's structure. This form of hierarchical spatial model has been implemented in a computer program called SPATIAL-LAYOUT, details of which are given in (Howarth and Buxton 1992) and Howarth (1994, 1998), which describes: how a spatial decomposition can be used to form a hierarchical database that holds the spatial information about the static environment; how an interface can be formed between the database and the rest of the runtime system that enables interpretation of dynamic object data; and how all the regular and irregular implementations would work, since they support the cellular representation we introduced above, although, some approaches are more attractive than others in terms of storage, runtime performance or additional representational power.

4. Conclusion

This survey gives some indication of the broad range of AI techniques that are connected to spatial representation and which can assist with the understanding and interpretation of those various unfolding activities that take place in a typical wide-area surveillance application. These spatial models are likely to play a much more important role in advanced visual surveillance once it becomes commonplace to recover depth information such as the ground-plane map shown in Figure 1 and the vehicle 3D positions and poses of Figure 5. Techniques for reasoning about the observed behaviour of these dynamic objects is surveyed in a related paper (Howarth 1995).

In this paper, we have described how spatial representation is a common theme running through various computational approaches that have been used for different objectives in AI research. We have looked in detail at the mathematical foundation provided by space filling cells; at the use of polyhedra to specify the geometry of the 2D ground-plane; and at how to map a polyhedral specification to a cellular representation.

Acknowledgements

This paper has been based on Howarth (1994). I thank Hilary Buxton who supervised this work. The research was funded, at various stages, by the EPSRC grant GR/K08772, the ESPRIT II project P2152 VIEWS, and a SERC CASE award with GEC-Marconi Research Centre. Figure 5 is from Figure 10 on page 27 of Corrall and Hill (1992). I thank Dave Corral and Graham Hill for giving me permission to reprint this figure. Figure 9(b) is adapted from Figure 16 on page 240 of Latombe (1991). I thank Jean-Claude Latombe for giving me permission to adapt this figure.

Notes

1. The examples used in this paper are from the VIEWS project. For details see Corrall and Hill (1992), Buxton and Gong (1995), Howarth (1998), Tan et al. (1998), Worrall et al. (1991). The VIEWS project used two main exemplars: a road traffic application using data collected from the Bremer-Stern roundabout; and an airport holding area application concerned with the ground movements of aircraft and their supporting attendants.
2. Human motion analysis is discussed by Aggarwal and Cai (1999) and Gavrila (1999), and how it applies to surveillance is covered in (Collins et al. 2000, 2001) and (Regazzoni et al., 2001). Other example applications include: *rooms*, see Johnson and Hogg (2002), and offices, see Ayers and Shah (2001), with other indoor environments described by Regazzoni et al. (2001); *city centre streets*, see Johnson and Hogg (1996), Lee et al. (2000), Stauffer and Grimson (2000); *station platforms*, see Brémond and Ihonnat (1998); *car parks*, see Regazzoni et al. (2001) and Collins et al. (2000), are popular application domains, with the combination of stationary and slow moving vehicles, and the paths of people originating or terminating at a vehicle, or passing through the scene; *outdoor areas* as used by Rosin and Ellis (1991) for their intruder detection application; and *sporting events*, such as soccer, see Blocher and Schirra (1995), and American football, see Intille and Bobick (1995).
3. Collins et al. (2000, 2001) and Regazzoni et al. (2001) discuss examples that use road traffic applications. For example, Collins et al. (2001) describe how a car is tracked as it travels through the CMU campus and is passed between the fields of view of multiple, active cameras. Other example applications include: *roads*, see Howarth (1998), Koller et al. (1993), Nagel (1994), Tan et al. (1998), Worrall et al. (1991); *motorways/freeways*, see Beymer et al. (1997), Fernyhough et al. (2000), Huang and Russell (1998); *airport holding-areas*, see Corrall and Hill (1992); *airport taxiways*, see Howarth and Tsang (1998).
4. See Brady (1997), Buxton and Gong (1995), Draper et al. (1996) for how task relevant knowledge has been utilised in computer vision: Draper et al. describe how knowledge about the likely organisation of object features guides the interpretation of static images of buildings, say, if you see A then also look for B – they call

this "contextual indexing". Buxton and Gong describe various ways knowledge has been used to assist visual surveillance. Brady gives a more general overview, based around various example applications, with emphasis on what knowledge is mobilised to perform a perceptual task, how this knowledge is represented, and then how it is best utilised. Building in more knowledge is not a simple process, particularly for systems where the principal source of information is vision.

5. Recovering the ground-plane can provide additional contextual information as well as enormously simplify vehicle pose recovery for model matching (Worrall et al., 1994). Traffic surveillance systems can make use of straight road edges and parallel road markings to recover a projective transform or homography for mapping image coordinates to world coordinates. Worrall et al. (1994) describe an interactive camera calibration tool that is demonstrated on a motorway scene. Beymer et al. (1997) use a similar template approach to describe a three lane highway that they use for road traffic analysis. An alternative, where multiple cameras are available is to do as Lee et al. (2000) describe and recover the ground-plane across overlapping views by applying planar geometric constraints to the tracked movements of objects that are identified as being the same in different camera images and from this transform each of the multiple views of the scene's ground-plane to a combined overhead view. Also of interest is the description by Makris and Ellis (2002) of how they used Tsai's camera calibration method to display their path learning results in a ground-plane projection.
6. Using some combination of multiple, static or active, cameras (see, Collins et al. 2001) introduces additional problems such as: how best to locate the cameras in their environment (also known as the "art gallery problem", see Marengoni et al. 2000; Tarabanis et al. 1995); how to take advantage of overlapping fields of view (see Chang et al. 2000; Lee et al. 2000); and other issues – for example, Hall and Llinas (1997) discuss multisensor data fusion and, while not including multi-camera issues, they do cover military and civilian applications and the use of AI approaches. Also, in relation to Endnote 5, Bradshaw et al. (1997, p. 221, 232) explain how an active camera can be used to recover the 2D ground plane by "calibrating the homography between the frontal plane" – which is "a plane perpendicular to the resting gaze direction and an arbitrary distance in front of the rotation centre of the camera" – and the "scene [ground] plane using plane to plane correspondences of at least four points or at least four lines". Line calibration is more complex since it involves tracing each line actively by the camera fovea. To illustrate the calibration technique they use trajectories created by people entering or leaving a building to reconstruct the scene trajectory of each target during tracking. They use this to demonstrate how an active camera platform, in a real-time implementation, can recover trajectories, typically in the ground plane of the scene, that can then be used to predict, using a Kalman filter, over the delays in the visual feedback loop.
7. Mobile platforms include: pedestrian detection from cameras attached to the sides of a bus or the top of a minivan (Zhao and Thorpe 2000); detecting the positions and activity of other road users in front of a moving car (Ferryman et al. 2000); doing aerial traffic/vehicle surveillance from an autonomous helicopter (the WITAS UAV project: Coradeschi et al. 1999; Doherty et al. 2000); and others that use a small unmanned plane are described by Collins et al. (2000) and Regazzoni et al. (2001). Related issues are discussed by Mayol et al. (2002) who have

developed a shoulder mounted wearable visual robot – here the mobile platform is the person wearing the camera, who interacts with objects in the world. This illustrates the three components (events; attention; environment) from the introduction of this paper, each having their own frame-of-reference – see Herskovits (1986, pp. 157–159); Howarth (1998, pp. 20–21).

8. For a discussion of how background pixels are identified see the comparison by Toyama et al. (1999) of their WALLFLOWER algorithm against eight other techniques for background subtraction.
9. This concept of the world being its own best model is explained by Brooks (1991, p. 583) who while providing background to this makes the following observations (Brooks 1991, p. 579):

> "The fundamental issue is that Artificial Intelligence and Computer Vision have made an assumption that the purpose of vision is to reconstruct the static external world (for dynamic worlds it is just supposed to do it often and quickly) as a three dimensional world model. I do not believe that this is possible with the generality that is usually assumed. Furthermore I do not think it is necessary, nor do I think that it is what human vision does."

This last point is demonstrated by the "change blindness" experiments of O'Regan and Noë (2001) that suggest rather than having a detailed and rich internal representation of the outside world, instead what we have is rather sparse, supplemented by immediate access to the information in the world, rather like accessing some external memory store. While it can be very useful to take account of how human vision operates, particularly with regard to visual attention – see Tsotsos (2001) – surveillance applications are one instance where having a model of the "static external world" covering at least that part in the camera's field of view, is a good design option. For example, see Collins et al.'s (2001) description of their site model.

10. We discuss work that identifies special surveillance task related locations in the scene or image in Section 1.4. Also see where we discuss the work of Ng and Gong (2002b) in Section 2.5. For more details about visual attention and how it can be used for motion understanding see Tsotsos' (2001) overview that discusses event representation and lists those assumptions often made to reduce the need for visual attention (such as the fixed camera assumption made in Section 1.2 – although you still have the problem of searching for interesting content within any image sequence from the constant field of view).
11. The various spaces (*visible proximal space*, *occluded proximal space*, etc.) are from (Howarth, 1994) and were selected to help delineate the perceptual space used by visual surveillance from the less relevant ones. Similar categorisations are described in Montello (1993), which contains an historical survey of scale distinctions before distilling insights from these into the classification shown in Table 2. Montello identifies four classes of psychological spaces: *figural* (which can be subdivided into *pictorial* for 2D and *object* for 3D); *vista*; *environmental*; and *geographical*. These are distinguished by two functional properties (from p. 315):
    - "the basis of the *projective* size of the space relative to the human body, not its actual or apparent absolute size"; and
    - how much locomotion around a space is necessary to apprehend it.

Table 2. Montello's four classes of psychological spaces. See Endnote 11.

	Projective size	Locomotion	Examples
Figural	smaller	none appreciable, one place	
—Pictural			small flat
—Object			small 3D
Vista	as large or larger	none appreciable, single place	single rooms, town squares, small valleys, horizons
Environmental	projectively larger	considerable over significant periods of time	buildings, neighbourhoods, cities
Geographical	projectively much larger than body	cannot be directly apprehended – learned via symbolic representations	states, countries, the solar system (maps, models – maps are instances of pictorial space)

And are represented, in Table 2, by the columns "Projective size" and "Locomotion" respectively. The third column provides illustrative examples. Of these four classes, the most pertinent to visual surveillance is the well named *vista space*, which corresponds to the *visible space* described in Section 1.5, and also has the connotation of looking at a landscape from some well chosen vantage point – a view through or between intervening objects such as an avenue of trees or buildings – which fits well with the geographic forms he is interested in. Both Montello's classification and the one from Section 1.5 describe a hierarchy of spatial scale. Selecting the most appropriate scale is important and is discussed further in Montello (1993, 2001a) which cover issues linking the scale of a geographic phenomenon with its analysis and representation. Other aspects of how spatial cognition is treated by geographers are described in a related paper by Montello (2001b), which gives both historic background and overlapping issues of interest, such as, the structures and processes of spatial knowledge (see Sections 2.1 and 2.2) and the communication via language of spatial information (see Section 2.6). And indicates the broader concerns of spatial cognition and how they are connected to the focus of visual surveillance covered in this survey. This shared interest in how best to represent and understand perceptions of the physical world is illustrated by Montello's introductory statement (2001b, p. 14771):

> "Spatial cognition concerns the study of knowledge and beliefs about spatial properties of objects and events in the world. Cognition is about knowledge: its acquisition, storage and retrieval, manipulation, and use by humans, nonhuman animals, and intelligent machines. Broadly construed, cognitive systems include sensation and perception, thinking, imagery, memory, learning, language,

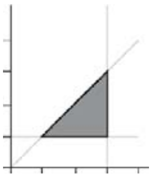
> reasoning, and problem solving. In humans, cognitive structures and processes are part of the mind, which emerges from a brain and nervous system inside of a body that exists in a social and physical world. Spatial properties include location, size, distance, direction, separation and connection, shape, pattern, and movement."

Which brings us back to the selection of the most appropriate spatial form for visual surveillance, in Section 1.5, where the emphasis is on how a space is perceived rather than projective size or locomotion. While projective size and locomotion are related to perception – all three concern how we experience space and are likely to be important to work on visual surveillance that uses mobile platforms such as those discussed in Endnote 7 – obviously locomotion is not going to be a concern for implementions that use a single static camera. And while perceived spatial form and projective size are useful and are similar, they differ in how they treat length (and depth and distance), for example, in Section 1.5 Euclidean space is used to build various qualitative spheres of perception around the observer, where as Montello (1993, p. 315) seems to be treating his four classes with a representational equivalence (perhaps in some projective or affine space) so that: "Large spaces viewed from a distance effectively become small spaces ... "; because to Montello "The most critical functional consequences of the relative projective size of a space are the means by which it may be apprehended and its cognitive treatment by the mind." This illustrates how purpose affects spatial representation and analysis, with Montello's bias towards the cartographic applications of the geographer, contrasting against the bias towards visual surveillance expressed in this paper.

12. Figure 2 shows the petrol station from which video sequences were collected on the VIEWS project, and the tram lines that pass through the middle of the roundabout – see Howarth and Buxton (1992) for another image that includes a tram in the shot. The road map, being just concerned with information the motorist needs, excludes unnecessary details. On VIEWS a more detailed road map, perhaps the original construction map itself, was used to input the various structural elements: road surface, pedestrian walkways, cycle ways, islands, tramway. Part of this is shown in Figure 1 and described in Howarth and Buxton (1992). And, although this map was accurate in most features, the configuration of the junction nearest the camera did not match the road markings and other features in this part of the video image. Illustrating the difficulty of obtaining and maintaining accurate cartographic data, which may be necessary for surveillance applications, particularly if the position of image features is being used to recover the ground-plane.
13. CONDOR (Strat and Fischler, 1991) makes extensive use of context, not just in how the image data is modelled but also in how the knowledge-based program is organised and how the condition-action rules are written. Another example of this is given by Draper et al. (1996) – see Endnote 4.
14. Hutchins (1983) describes how Micronesian navigators use an egocentric perspective of the oceanic environment through which they are navigating. Compared with the Western use of map and compass the Micronesian navigators place a very different interpretation on the scene primarily based on ocean swells, the stars and the positions of reference islands. A similar description is given by Oatley (1977). Also, some aspects of this difference in perspective are illustrated in

Figure 1 which shows an example camera image and a corresponding ground-plane projection.

15. A semi-algebraic set of polynomial equalities or inequalities can be used to define the interior or exterior of an object – see Canny (1988, p. 85), Latombe (1991, p. 146), Yap (1987, p. 124), Sections 3.1.5 and 3.3.2. Useful introductions to semi-algebraic sets are given by Dumortier et al. (1997) and Paredaens and Kuijpers (1998) explaining how a closed triangle in a plane can be described by $\{(x, y) \mid x \leq 3 \wedge y \geq 1 \wedge x \geq y\}$ with the open triangle being $\{(x, y) \mid x < 3 \wedge y > 1 \wedge x > y\}$ and its boundary being $\{(x, y) \mid (x = 3 \wedge 1 \leq y \leq 3) \vee (y = 1 \wedge 1 \leq x \leq 3) \vee (x = y \wedge 1 \leq x \leq 3)\}$.



16. For example, the more restrictive form of semi-algebraic sets, called algebraic sets, can be used to define polynomial ideals – Hoffmann (1989, p. 257) says these are "sets of polynomials that describe elementary geometric objects". These polynomial ideals can be manipulated by the Gröbner bases algorithm as described by Buchberger (1988), Cox et al. (1992) for various geometrical applications including constructive solid geometry (Hoffmann 1989, pp. 257–301) and reasoning about 2D shapes (Schweitzer and Straach 1995).
17. Cohn and Hazarika (2001) place the region connection calculus (RCC-8) of Randell and Cohn in the context of other qualitative spatial reasoning techniques, explaining how, this elegant description of topological arrangements between two regions and the allowed transitions between these arrangements, has been developed further. For example, Gerevini and Renz (2002) present RCC-7, a subalgebra of RCC-8, that can be used for applications where spatial regions cannot partially overlap.
18. Both Herskovits (1986) and Talmy (1983) give examples of how "geometric" properties are used in language, how objects are characterised almost solely by qualitative or topological properties, how localising an object to determine its location can involve dividing space into subregions, and how basic spatial discriminations are made by language. Also related issues of how spatial language expresses object identification and object location are discussed by Landau and Jackendorff (1993) and Hayward and Tarr (1995). For more details on language and space see Bloom et al. (1996) and for a broader cognitive perspective see Barwise (1981), Mark and Frank (1991) and McKevitt (1996).
19. It could be useful to have an automatically generated linguistic description giving the arrangement of objects in a scene, by assigning spatial relationships (left-of, right-of, above, below, ...) to the labelled entities on a 2D segmented image of this scene. This is done by Keller and Wang (2000), who use a fuzzy rule-based approach. Alternative fuzzy-logic approaches are proposed by Matsakis and Wendling (1999), who use a histogram of forces, and Jan and Hsueh (2000), who use influence zones to find the skeleton line that is equidistant between two regions.
20. See Tsotsos (2001) description of Badler's pioneering work that used motion verbs to produce conceptual descriptions from dynamic scenarios depicting the 3D world. Also see the equally important use of motion verbs by Neumann (1983, 1989) and Novak and Neumann (1986) in their NAOS system to generate textual descriptions of traffic scenes. These ideas have been developed further by Herzog and Wazinski (1994), Mann et al. (1997), Nagel (1994) and Rao et al. (2002). For example, Rao et al. show how a trajectory, from tracking a person's hand movements when performing a given task, can be segmented into "intervals" of activ-

ity and places of state change (see Fleck 1996, p. 21) they call "dynamic instants" by identifying peaks in the value of the trajectory's spatio-temporal curvature. To demonstrate the validity of the segmentation they hand-label motion verbs to the intervals and use background reference objects and reference frames to explain and justify the labelling. Illustrating the connection between perceived activity and natural language and also how, as in Figure 4, the meaning attached to the trajectory is afforded by environmental context.

21. Superquadrics are popular shape primitives. For example, Chella et al. (2001) use them as constructive solid geometry primitives (see, for example, Hoffmann 1989) and as instances in a conceptual space of simple shapes. For further details, together with an extensive survey describing other deformable surfaces, see Montagnat et al. (2001).
22. For example, Scaife and Rogers (1996) identify the presence of a "resemblance fallacy" that holds between an external graphical representation and its internal representation, which they say is due to the assumption or intuition that these have the same characteristics, when there is no well-articulated theory that shows this is so. Although see Kosslyn (1994).
23. See Endnote 1.
24. The *regular* in regular cell complex does not refer to the pattern formed by the cells. Instead it describes a complex where "all its cells are regular cells" (Fritsch and Piccinini 1990, p. 23).
25. Felix Hausdorff (1868–1942) gave the definition of the topological space that is today known as "a Hausdorff space" in Hausdorff (1914).
26. Lemma 5 requires a different interpretation to the cellular representation of the sphere, such as Bredon's example "8.4" (1993, p. 197) which has two 0-cells, two 1-cells, two 2-cells, ..., two n-cells. Other examples include: the Collins decomposition given in Figure 9(b), and the simplicial representation in Figure 7. And also the square cell and triangular cell implementations discussed in Howarth (1994).
27. This approach is related to constructive solid geometry (see Hoffmann 1989, pp. 62–63 and Section 2.4). The "pile of sugar cubes" example is from Koenderink (1990, p. 48, 609). Other names for this regular volumetric model made out of cubes are "occupancy array" and "voxel" – volumetric element. This has been used in computer vision (see Sonka et al. 1993, p. 380; Seitz and Dyer 1999; Sato et al. 2002) and in medical imaging (see Udupa 1983; Ackerman 1998) and by Glasgow and Papadias (1992, p. 370) to model the atomic components of molecules. Also Mitchell (1990, pp. 40–41) describes how Bemis (1936) used voxels (4" cubes) in the architectural design of prefabricated buildings. Alternatives to cubes include tetrahedrons (see Jung and Lee 1993). And other forms of space-filling polyhedra are described by Gasson (1983, pp. 249–254) and Lord and Wilson (1984, pp. 164–165).
28. A regular polygon has all angles equal and all sides equal. Grünbaum and Shephard (1987) and Lord and Wilson (1984, pp. 152–153) describe how regular patterns in the plane can also be formed by other shapes like parallelograms (rectangle, rhombus) and others that form tiling patterns. Also see Bell et al. (1983), and the pseudo hexagonal tessellation described by Fleck (1991).
29. See Endnotes 15 and 16, as well as, Sections 2.3 and 2.4, where semi-algebraic sets were discussed as part of robot motion planning and computational geometry.

References

Ackerman, M. J. (1998). The Visible Human Project. *Proceedings of the IEEE* **86**(3): 504–511.

Adams, N. J. & Williams, C. K. (2002). Dynamic trees: Learning to Model Outdoor Scenes. In Heyden, A., Sparr, G., Nielsen, M., & Johansen, P. (eds.) *Proceedings of the Seventh ECCV Conference*, *Vol. IV.* Copenhagen, Denmark, 82–96, Lecture Notes in Computer Science 2353, Springer-Verlag, Berlin.

Aggarwal, J. & Cai, Q. (1999). Human Motion Analysis: A Review. *Computer Vision and Image Understanding* **73**(3): 428–440.

Alexandroff, P. (1961). *Elementary Concepts in Topology*. Dover Publications, Inc., New York. An English translation of *Einfachste Grundbegriffe der Topologie*, Julius Springer: Berlin, 1932.

Asher, N. & Vieu, L. (1995). Toward a Geometry of Common Sense: A Semantics and Complete Axiomatization of Mereotopology. In *Proceedings of the Fourteenth IJ-CAI Conference*, 846–852. Montréal, Canada.

Ayers, D. & Shah, M. (2001). Monitoring Human Behavior from Video Taken in an Office Environment. *Image and Vision Computing* **19**(12): 833–846.

Badler, N. I. & Smoliar, S. W. (1979). Digital Representation of Human Movement. *Computing Surveys* **11**(1): 19–38.

Barkowsky, T. & Freksa, C. (1997). Cognitive Requirements on Making and Interpreting Maps. In Hirtle, S. C. & Frank, A. U. (eds.) *Spatial Information Theory: A Theoretical Basis for GIS*, 347–361. Lecture Notes in Computer Science 1329, Springer-Verlag.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences* **22**(4): 577–660.

Barwise, J. (1981). Scenes and Other Situations. *The Journal of Philosophy* **LXXVIII**(7): 369–397.

Bell, S. B., Diaz, B., Holroyd, F. C. & Jackson, M. (1983). Spatially Referenced Methods of Processing Raster and Vector Data. *Image and Vision Computing* **1**(4): 211–220.

Bell, S. B., Holroyd, F. C. & Mason, D. C. (1989). A Digital Geometry for Hexagonal Pixels. *Image and Vision Computing* **7**(3): 194–204.

Bemis, A. F. (1936). *The Evolving House, Volume III: Rational Design*. The Technology Press, M.I.T.: Cambridge, Mass. *The Evolving House*, 1933–1936, in three volumes.

Beymer, D., McLauchlan, P. F., Coifman, B. & Malik J. (1997). A Real-time Computer Vision System for Measuring Traffic Parameters. In *Proceedings IEEE Computer Vision and Pattern Recognition*. San Juan, Puerto Rico, pp. 495–501, IEEE Press. Similar version in *Transportation Research Part C: Emerging Technologies*, **6**(4): 271–288, August 1998.

Blocher, A. & Schirra, J. R. (1995).Optional Deep Case Filling and Focus Control with Mental Images: ANTLIMA-KOREF. In *Proceedings of the Fourteenth IJCAI Conference*, 417–423. Montréal, Canada: Lidee.

Blocher, A. & Stopp, E. (1998). Time-Dependent Generation of Minimal Sets of Spatial Descriptions. In Olivier, P. & Gapp, K.-P. (eds.) *Representation and Processing of Spatial Expressions*, 57–72. Mahwah, NJ: Lawrence Erlbaum Associates.

Bloom, P., Peterson, M. A., Nadel, L. & Garrett, M. F. (eds.) (1996). *Language and Space*. The MIT Press: Cambridge, MA.

Bobick, A. F. & Davis, J. W. (2001). The Recognition of Human Movement using Temporal Templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23**(3): 257–267.

Brachman, R. J. & Levesque, H. J. (eds.) (1985). *Readings in Knowledge Representation*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Bradshaw, K. J., Reid, I. D. & Murray, D. W. (1997). The Active Recovery of 3D Motion Trajectories and Their Use in Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(3): 219–234.

Brady, M. (1997). The Forms of Knowledge Mobilized in Some Machine Vision Systems. In Anderson, J. Barlow, H. & Gregory, R. (eds.) *Knowledge-Based Vision in Man and Machine*, Vol. 352 (number 1358). 1241–1248. Philosophical Transactions of the Royal Society of London SERIES B: Biological Sciences.

Bredon, G. E. (1993). *Topology and Geometry*. Springer-Verlag, Berlin. Graduate texts in mathematics 139.

Brémond, F. & Thonnat, M. (1998). Issues of Representing Context Illustrated by Video-Surveillance Applications. *International Journal of Human–Computer Studies* **48**(3): 375–391. Special issue on Using Context in Applications.

Brooks, R. A. (1991). Intelligence Without Reason. In *Proceedings of the Twelfth IJ-CAI Conference*, 569–595. Darling Harbour, Sydney, Australia.

Buchberger, B. (1988). Applications of Gröbner bases in Non-Linear Computational Geometry. In Rice, J. R. (ed.) *Mathematical Aspects of Scientific Software, IMA Volume in Mathematics and its Applications 14*. Springer-Verlag, Berlin. Also in Kapur and Mundy (1989), 413–446.

Buxton, H. & Gong, S. G. (1995). Visual Surveillance in a Dynamic and Uncertain World. *Artificial Intelligence* **78**(1–2): 431–459.

Canny, J. F. (1988). *The Complexity of Robot Motion Planning*. The MIT Press: Cambridge, MA.

Chang, T.-H., Gong, S. & Ong, E.-J. (2000). Tracking Multiple People Under Occlusion using Multiple Cameras. In *Proceedings of the Eleventh British Machine Vision Conference* 566–575. *Volume II*. Bristol, UK, BMVA Press. On-line at http://www.bmva.ac.uk/bmvc/2000/papers/p57.pdf.

Chapman, D. (1991). *Vision, Instruction and Action*. The MIT Press: Cambridge, MA.

Chazelle, B. & Dobkin, D. (1985) Optimal Convex Decompositions. In Toussaint, G. (ed.) *Computational Geometry*, 63–133, North-Holland Publ. Co. (Elsevier), Amsterdam.

Chella, A., Frixione, M. & Gaglio, S. (2001). Conceptual Spaces for Computer Vision Representations. *Artificial Intelligence Review* **16**(2): 137–152.

Chen, S.-S. (ed.) (1990). *Advances in Spatial Reasoning*. Ablex Publishing Corporation: Norwood, New Jersey. Two volumes.

Clarke, B. L. (1981). A Calculus of Individuals Based on "Connection". *Notre Dame Journal of Formal Logic* **22**(3): 204–218.

Clarke, B. L. (1985). Individuals and Points. *Notre Dame Journal of Formal Logic* **26**(1): 61–75.

Coenen, F., Beattie, B., Shave, M., Bench-Capon, T. & Diaz, B. (1998). Spatial Reasoning Using the Quad Tesseral Representation. *Artificial Intelligence Review* **12**(4): 321–343.

Cohn, A. G. & Hazarika, S. M. (2001). Qualitative Spatial Representation and Reasoning: an Overview. *Fundamenta Informaticae* **46**(1–2): 1–29.

Collins, A. & Smith, E. E. (eds.) (1988). *Readings in Cognitive Science*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Collins, G. E. (1975). Qualifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition. In Goos, G. & Hartmanis, J. (eds.) *Automata Theory and Formal Languages: Proceedings of the Second GI Conference*, 134–183. Kaiserslautern, Lecture Notes in Computer Science 33, Springer-Verlag.

Collins, R. T., Lipton, A. J., Fujiyoshi, H. & Kanade, T. (2001). Algorithms for Cooperative Multisensor Surveillance. *Proceedings of the IEEE* **89**(10): 1456–1477.

Collins, R. T., Lipton, A. J. & Kanade T. (2000). Introduction to the Special Section on Video Surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8): 745–746.

Coradeschi, S., Karlsson, L. & Nordberg, K. (1999). Integration of Vision and Decision-Making in an autonomous Airborne Vehicle for Traffic Surveillance. In Christensen, H. I. (ed.) *Proceedings of the First International Conference on Computer Vision Systems (ICVS)*, 216–230, Las Palmas de Gran Canaria, Spain, Lecture Notes in Computer Science 1542, Springer-Verlag.

Corrall, D. R. & Hill, A. G. (1992). Visual Surveillance. *GEC Review* **8**(1): 15–27.

Cox, D. A., Little, J. B. & O'Shea, D. (1992). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, Berlin. Undergraduate Texts in Mathematics. Second edition, 1996.

Cui, Z., Cohn, A. & Randell, D. (1992). Qualitative Simulation Based on a Logical Formalism of Space and Time. In *Proceedings of the Tenth AAAI Conference*, 679–684. San Jose, California, The AAAI Press/The MIT Press.

Davis, E. (1986). *Representing and Acquiring Geographic Knowledge*. Pitman Publishing: London, England.

Davis, E. (1988). Inferring Ignorance from the Locality of Visual Perception. In *Proceedings of the Seventh AAAI Conference*, 786–790. Saint Paul: Minnesota.

Davis, E. (1989). Solution to a Paradox of Perception with Limited Acuity. In Brachman, R., Levesque, H. & Reiter, R. (eds.) *KR '89: Principles of Knowledge Representation and Reasoning*. Proceedings of the First Conference, Toronto, Ontario, Canada, 79–82, Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Davis, E. (1990). *Representations of Commonsense Knowledge*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

de Kleer, J. (1977). Multiple Representation of Knowledge in a Mechanics Problem Solver. In *Proceedings of the Fifth IJCAI Conference*, 299–304. Cambridge, Massachusetts.

Dodson, C. T. (1974). Hazy Spaces and Fuzzy Spaces. *The Bulletin of the London Mathematical Society* **6**: 191–197.

Doherty, P., Granlund, G., Kuchcinski, K., Sandewall, E., Nordberg, K., Skarman, E. & Wiklund, J. (2000). The WITAS Unmanned Aerial Vehicle Project. In Horn W. (ed.) *ECAI 2000. Proceedings of the Fourteenth European Conference on Artificial Intelligence*. Berlin, pp. 747–755, Amsterdam: IOS Press.

Dorst, L. (1998). Analyzing the Behaviors of a Car: A Study in Abstraction of Goal-Directed Motions. *IEEE Transactions on Systems, Man, and Cybernetics* **28**(6): 811–822.

Downs, R. M. & Stea, D. (eds.) (1973). *Image and Environment: Cognitive Mapping and Spatial Behaviour*. Aldine Publishing Company, Chicago.

Draper, B. A., Hanson, A. R. & Riseman E. M. (1996). Knowledge-Directed Vision: Control, Learning, and Integration. *Proceedings of the IEEE* **84**(11): 1625–1637.

Dumortier, F., Gyssens, M. Vandeurzen, L. & Gucht, D. V. (1997). On the Decidability of Semi-linearity for Semi-algebraic Sets and its Implications for

Spatial Databases (extended abstract)'. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 97)*. Tucson, Arizona USA, pp. 68–77, ACM. On-line at http://www.acm.org/ and http://www.luc.ac.be/research/groups/theocomp/papers.

Durrant-Whyte, H. F. (1988). Uncertain Geometry in Robotics. *IEEE Journal of Robotics and Automation* **4**(1): 23–31. Also, different version in Kapur and Mundy (1989), pages 447–481.

Duygulu, P., Barnard, K. de Freitas, J. & Forsyth, D. (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In Heyden, A., Sparr, G., Nielsen, M. & Johansen, P. (eds.) *Proceedings of the Seventh ECCV Conference*, Vol. IV. Copenhagen, Denmark, 97–112. Lecture Notes in Computer Science 2353, Springer-Verlag.

El Gindy, H. & Toussaint, G. T. (1989). On Geodesic Properties of Polygons Relevant to Linear Time Triangulation. *The Visual Computer* **5**: 68–74.

Feng, X., Williams, C. K. & Felderhof, S. N. (2002). Combining Belief Networks and Neural Networks for Scene Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(4): 467–483.

Fernyhough, J. H., Cohn, A. G. & Hogg, D. C. (1996). Generation of Semantic Regions from Image Sequences. In Buxton, B. & Cipolla, R. (eds.) *Proceedings of the Fourth European Conference on Computer Vision*, 475–484. Vol. II. Cambridge, UK., Lecture Notes in Computer Science 1065, Springer-Verlag.

Fernyhough, J. H., Cohn, A. G. & Hogg, D. C. (2000). Constructing Qualitative Event Models Automatically from Video Input. *Image and Vision Computing* **18**(2): 81–103.

Ferryman, J. M., Maybank, S. J. & Worrall, A. D. (2000). Visual Surveillance for Moving Vehicles. *International Journal of Computer Vision* **37**(2): 187–197.

Fischler, M. A., & Firschein, O. (eds.) (1987). *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Fleck, M. M. (1988a). Boundaries and Topological Algorithms. Ph.D. thesis, MIT AI Lab. AI-TR 1065.

Fleck, M. M. (1988b). Representing Space for Practical Reasoning. *Image and Vision Computing* **6**(2): 75–86.

Fleck, M. M. (1991). A Topological Stereo Matcher. *International Journal of Computer Vision* **6**(3): 197–226.

Fleck, M. M. (1996). The Topology of Boundaries. *Artificial Intelligence* **80**(1): 1–27.

Forbus, K. D. (1983). Qualitative Reasoning about Space and Motion. In Dedre, G. & Stevens A. (eds.) *Mental Models*, 53–73. Lawrence Erlbaum Associates: Hillsdale, NJ.

Forbus, K. D., Nielsen, P. & Faltings, B. (1987). Qualitative Kinematics: A Framework. In *Proceedings of the Tenth IJCAI Conference*. Milan, Italy, 430–436. Also in Weld and de Kleer (1990), 562–567.

Forbus, K. D., Nielsen, P. & Faltings, B. (1991). Qualitative Spatial Reasoning: the CLOCK project. *Artificial Intelligence* **51**: 417–471.

Fortune, S. (1987). Sweepline Algorithm for Voronoi Diagrams. *Algorithmica* **2**(2): 153–174. Also in *Proceedings of the Second ACM Symposium on Computational Geometry*, Yorktown Heights: New York, 313–319, 1986.

Freksa, C. (1992). Using Orientation Information for Qualitative Reasoning. In Frank, A., Campari, I. & Formentini, U. (eds.) *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. 162–178. Lecture Notes in Computer Science 639, Springer Verlag.

Fritsch, R. & Piccinini, R. A. (1990). *Cellular Structures in Topology*. Cambridge University Press, Cambridge, England.

Funt, B. V. (1980). Problem-solving with Diagrammatic Representations. *Artificial Intelligence* **13**(3): 201–230. Also in Fischler and Firschein (1987), 456–470 and Brachman and Levesque (1985), 441–456.

Funt, B. V. (1983). Analogical Models of Reasoning and Process Modeling. *IEEE Computer Magazine* **16**(10): 99–104.

Gallistel, C. R. (1990). *The Organization of Learning*. The MIT Press: Cambridge, MA.

Gardina, F. & Meltzera, B. (1989). Analogical Representations of Naive Physics. *Artificial Intelligence* **38**(2): 139–159.

Gargantini, I. (1982). An Efficient Way to Represent Quadtrees. *Communications of the ACM* **25**(12): 905–910.

Gasson, P. C. (1983). *Geometry of Spatial Forms: Analysis, Synthesis, Concept Formulation and Space Vision for CAD*. Ellis Horwood Ltd, Chichester, West Sussex, England.

Gavrila, D. M. (1999). The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding* **73**(1): 82–98.

Gelsey, A. (1987). Automated Reasoning about Machine Geometry and Kinematics. In *Proceedings of the Third IEEE Conference on AI Applications*, 182–187. Orlando, Florida, Also in Weld and de Kleer (1990), 580–591.

Gerevini, A. & Renz, J. (2002). Combining Topological and Size Information for Spatial Reasoning. *Artificial Intelligence* **137**(1–2): 1–42.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin Company.

Glasgow, J. & Papadias, D. (1992). Computational Imagery. *Cognitive Science* **16**(3): 355–394.

Glasgow, J. I. (1993). The Imagery Debate Revisited: A Computational Perspective. *Computational Intelligence* **9**(4): 309–333. A lead article with response pages 424–435.

Gold, C. M. (1990). Spatial Data Structures: The Extension from One to Two Dimensions. In Pau L. (ed.) *Mapping and Spatial Modelling for Navigation*. NATO ASI series F: computer and systems science, vol. 65, Springer-Verlag, 11–39.

Gold, C. M. (1992). The Meaning of "Neighbour". In Frank, A., Campari, I. & Formentini U. (eds.) *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space* 220–235. Lecture Notes in Computer Science 639, Springer-Verlag.

Green, D. H. (1983). The Decomposition of Polygons into Convex Parts. In Preparatra, F. P. (ed.) *Advances in Computing Research: volume 1*, 235–259. JAI Press, Greenwich, Connecticut.

Grünbaum, B. & Shephard, G. C. (1987). *Tiling and Patterns*. W.H. Freeman & Co., New York.

Guibas, L. J. & Stolfi, J. (1985). Primitives for the Manipulation of General Subdivisions and the Computation of Voronoi Diagrams. *ACM Transactions on Graphics* **4**(2): 74–123. Also in *Proceedings of the Fifteenth ACM Symposium on the Theory of Computing (STOC)*, Boston, Massachusetts, 221–234, April 1983.

Günther, O. (1988). *Efficient Structures for Geometric Data Management*. Lecture Notes in Computer Science 337, Springer-Verlag.

Günther, O. & Bilmes, J. (1991). 'Tree-based methods for spatial databases: implementations and performance evaluation'. *IEEE Transactions on Knowledge and Data Engineering* **3**(3): 342–355.

Hall, D. L. & Llinas, J. (1997). 'An introduction to multisensor data fusion'. *Proceedings of the IEEE* **85**(1): 6–23.

Hausdorff, F. (1914). *Grundzüge der Mengenlehre (Essentials of Set Theory)*. Verlag von Veit & Comp.: Leipzig, Germany. Reprinted by Chelsea Publishing Company: New York, 1949 and also by Springer-Verlag: Heidelberg, 2002 as part of their *Gesammelte Werke Band II (Collected Works Volume Two)*.

Hayes, P. (1985). 'The second naive physics manifesto'. In Hobbs, J. R. & Moore, R. C. (eds.) *Formal Theories of the Commonsense World*, 1–36, Ablex Publishing Corporation, Norwood, New Jersey.

Hayward, W. G. & Tarr, M. J. (1995). 'Spatial language and spatial representation'. *Cognition* **55**(1): 39–84.

Heider, F. & Simmel, M. (1944). 'An experimental study of apparent behavior'. *American Journal of Psychology* **57**: 243–259.

Hernández, D. (1994). *Qualitative Representation of Spatial Knowledge*. Lecture Notes in Artificial Intelligence 804, Springer-Verlag.

Herring, J. R. (1991). 'The mathematical modeling of spatial and non-spatial information in geographic information systems'. In Mark, D. & Frank, A. (eds.) *Cognitive and Linguistic Aspects of Geographic Space*, 313–350, Kluwer Academic Publ., Dordrecht and Boston.

Herskovits, A. (1986). *Language and Spatial Cognition: An interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge, England.

Herzog, G. & Wazinski, P. (1994). 'VIsual TRAnslator: linking perceptions and natural language descriptions'. *Artificial Intelligence Review* **8**(2–3). 175–187.

Hobby, J. D. (1993). 'Generating automatically tuned bitmaps from outlines'. *Journal of the ACM* **40**(1): 48–94.

Hoffmann, C. M. (1989). *Geometric and Solid Modeling: An Introduction*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Horn, B. K. P. (1986). *Robot Vision*. The MIT Press, Cambridge, MA.

Howarth, R. J. (1994). *Spatial Representation, Reasoning and Control for a Surveillance System*. Ph.D. thesis, Queen Mary and Westfield College, University of London.

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

Howarth, R. J. (1998). 'Interpreting a dynamic and uncertain world: task-based control'. *Artificial Intelligence* **100**(1–2): 5–85.

Howarth, R. J. & Buxton, H. (1992). 'An analogical representation of space and time'. *Image and Vision Computing* **10**(7): 467–478.

Howarth, R. J. & Tsang, E. P. K. (1998). 'Spatio-temporal conflict detection and resolution'. *Constraints* **3**(4): 343–361.

Huang, T. & Russell, S. (1998). 'Object identification: a Bayesian analysis with application to traffic surveillance'. *Artificial Intelligence* **103**(1–2): 77–93.

Hutchins, E. (1983). 'Understanding Micronesian Navigation'. In Dedre, G. & Stevens, A. (eds.): *Mental Models*, 191–225, Lawrence Erlbaum Associates, Hillsdale, NJ.

Intille, S. S. & Bobick, A. F. (1995). 'Closed-world tracking'. In *Proceedings of the Fifth International Conference on Computer Vision*, 672–678 Boston, Massachusetts, IEEE Computer Society Press.

Intille, S. S., Davis, J. W. & Bobick, A. F. (1997). 'Real-time closed-world tracking'. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 697–703, San Juan, Puerto Rico, IEEE Computer Society Press.

Jan, S. -R. & Hsueh, Y.-C. (2000). 'Primitive spatial relations based on SKIZ'. *Image and Vision Computing* **18**(8): 597–605.

Johnson, N. & Hogg, D. (1996). Learning the Distribution of Object Trajectories for Event Recognition. *Image and Vision Computing* **14**(8): 609–615. Also in Pycock, D. (Ed.), *Proceedings of the Sixth British Machine Vision Conference*, 583–592, Birmingham, England. BMVA Press, September 1995.

Johnson, N. & Hogg, D. (2002). Representation and Synthesis of Behaviour Using Gaussian Mixtures. *Image and Vision Computing* **20**(12): 889–894.

Johnson-Laird, P. N. & Wason, P. C. (1977). Introduction to Imagery and Internal Representation. In Johnson-Laird, P. N. & Wason, P. C. (eds.) *Thinking: Readings in Cognitive Science*, 523–531, Cambridge University Press, Cambridge, England.

Jung, Y. & Lee, K. (1993). Tetrahedron-based Octree Encoding for Automatic Mesh Generation. *Computer-Aided Design* **25**(3): 141–153.

Kapur, D. & Mundy, J. L. (eds.) (1989). *Geometric Reasoning*. The MIT Press, Cambridge, MA. Also published as *Artificial Intelligence*, **37**(1–3), 1988.

Kaufman, S. G. (1991). A Formal Theory of Spatial Reasoning. In Allen, J. Fikes, R. & Sandewall, E. (eds.) *KR '91: Principles of Knowledge Representation and Reasoning*. Proceedings of the Second Conference, Cambridge, Massachusetts, pp. 347–356, Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Keil, J. M. & Sack, J.-R. (1985). Minimum Decompositions of Polygonal Objects. In Toussaint, G. (ed.) *Computational Geometry*, 197–216, North-Holland Publ. Co. (Elsevier), Amsterdam.

Keller, J. M. & Wang, X. (2000). A Fuzzy Rule-based Approach to Scene Description Involving Spatial Relationships. *Computer Vision and Image Understanding* **80**(1): 21–41.

Koenderink, J. J. (1990). *Solid Shape*. The MIT Press, Cambridge, MA.

Koller, D., Daniilidis, K. & Nagel, H.-H. (1993). Model-based Object Tracking in Monocular Image Sequences of Road Traffic Scenes. *International Journal of Computer Vision* **10**(3): 257–281.

Kong, X., Everett, H. & Toussaint, G. (1990). The Graham Scan Triangulates Simple Polygons. *Pattern Recognition Letters* **11**: 713–716.

Kosslyn, S. M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. The MIT Press, Cambridge, MA.

Kramer, G. (1990). Solving Geometric Constraint Systems. In *Proceedings of the Eighth AAAI Conference*, 708–714, Boston, Massachusetts.

Kramer, G. A. (1992). *Solving Geometric Constraint Systems: A Case Study in Kinematics*. The MIT Press, Cambridge, MA.

Kuipers, B. (1978). Modeling Spatial Knowledge. *Cognitive Science* **2**: 129–153. Also in Chen (1990) vol 2, 171–198.

Kuipers, B. (2000). The Spatial Semantic Hierarchy. *Artificial Intelligence* **119**(1–2): 191–233.

Kuipers, B. & Levitt, T. (1988). Navigation and Mapping in Large-scale Space. *AI Magazine* **9**(2): 25–43. Also in Chen (1990) vol. 2, 207–251.

Landau, B. & Jackendorff, R. (1993). "What" and "Where" in Spatial Language and Spatial Cognition. *Behavioral and Brain Sciences* **16**(2): 217–265.

Lang, E., Carstensen, K.-U. & Simmons, G. (1991). *Modelling Spatial Knowledge on a Linguistic Basis: Theory, Prototype, Investigation*. Lecture Notes in Artificial Intelligence 481, Springer Verlag.

Latombe, J. (1991). *Robot Motion Planning*. Kluwer Academic Publ., Dordrecht and Boston.

Laumond, J.-P. (1993). Singularities and Topological Aspects in Nonholonomoc Motion Planning. In Li, Z. & Canny, J. (eds.) *Nonholonomic Motion Planning*, 149–199, Kluwer Academic Publ., Dordrecht and Boston.

Laurini, R. & Thompson: D. (1992) *Fundamentals of Spatial Information Systems*. Academic Press, London, England and San Diego, CA.

Lee, L., Romano, R. & Stein, G. P. (2000). 'Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **22**(8): 758–767.

Lefschetz, S. (1970). The Early Developments of Algebraic Topology. *Boletim Sociedade Brasileira De Mathematica* **1**: 1–48.

Lord, E. A. & Wilson, C. (1984). *The Mathematical Description of Shape and Form*. Ellis Horwood Ltd, Chichester, West Sussex, England.

Lozano-Pérez, T. (1983). Spatial Planning: A Configuration Space Approach. *IEEE Transaction on Computers* **32**: 108–120.

Lynch, K. (1960). *The Image of the City*. The MIT Press, Cambridge, MA.

Makris, D. & Ellis, T. (2002). 'Path Detection in Video Surveillance. *Image and Vision Computing* **20**(12): 895–903.

Mann, R., Jepson, A. & Siskind, J. M. (1997). The Computational Perception of Scene Dynamics. *Computer Vision and Image Understanding* **65**(2): 113–128.

Marengoni, M., Draper, B. Hanson, A. & Sitaraman, R. (2000). A System to Place Observers on a Polyhedral Terrain in Polynomial Time. *Image and Vision Computing* **18**(10): 773–780.

Mark, D. M. & Frank, A. U. (eds.) (1991). *Cognitive and Linguistic Aspects of Geographic Space*. Kluwer Academic Publ., Dordrecht and Boston. Proceedings of the NATO Advanced Studies Institute, Las Navas del Marqués, Spain, July 1990.

Marr, D. (1982). *Vision*. W.H. Freeman & Co., New York.

Massey, W. S. (1980). *Singular Homology Theory*. Springer-Verlag, Berlin.

Mataric, M. J. (1991). Navigating with a Rat Brain: A neurobiologically-inspired model for Robot Spatial Representation. In Meyer, J.-A. & Wilson, S. W. (eds.) *From Animals to Animats, the Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 169–175, Paris, France, The MIT Press, Cambridge, MA.

Matsakis, P. & Wendling, L. (1999). A New Way to Represent the Relative Position Between areal Objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **21**(7): 634–643.

Maunder, C. (1970). *Algebraic Topology*. Van Nostrad Reinhold Company, London, England.

Mayol, W. W., Tordoff, B. & Murray, D. W. (2002). Wearable visual robots. *Personal and Ubiquitous Computing* **6**(1): 37–48. Also in *Fourth International Symposium on Wearable Computers (ISWC'00)*, Atlanta, Georgia, p. 95, October 2000. Available from http://www.robots.ox.ac.uk/~bjt/Research/Wearable/.

McKevitt, P. (ed.) (1994–1996). *Integration of Natural Language and Vision Processing*. Kluwer Academic Publ., Dordrecht and Boston. Four volumes. *Volume I: Computational Models and Systems*, July 1995, reprinted from *Artificial Intelligence Review*, 8(2–3), 1994 and 8(5–6), 1994–5. *Volume II: Intelligent Multimedia*, October 1995, reprinted from *Artificial Intelligence Review*, 9(2–3), June 1995. *Volume III:*

Theory and Grounding Representations, June 1996, reprinted from *Artificial Intelligence Review*, *"Theory"* 9(4–5), October 1995 and *"Grounding Representations"* 10(1–2), April 1996. *Volume IV: Recent Advances*, September 1996, reprinted from *Artificial Intelligence Review*, 10(3–4), August 1996.

Mehlhorn, K. (1984). *Data Structures and Algorithms 3: Multi-dimensional Searching and Computational Geometry*. Springer-Verlag, Berlin. An English translation of *Effiziente Allgorithmen*, Stuttgart: B.G. Teubner publishers, 1977.

Mellor, J. (2003). Geometry and Texture from Thousands of Images. *International Journal of Computer Vision* **51**(1): 5–35.

Miller, G. A. & Johnson-Laird, P. N. (1976). *Language and Perception*. Harvard University Press, Cambridge, Massachusetts.

Minsky, M. (1975). A Framework for Representing Knowledge. In Wilson, P. H. (ed.) *The Psychology of Computer Vision*, 211–277, McGraw-Hill, New York. Also in Collins and Smith (1988), 156–189.

Mitchell, W. J. (1990). *The Logic of Architecture: Design, Computation, and Cognition*, The MIT Press, Cambridge, MA.

Mohnhaupt, M. & Neumann, B. (1991). Understanding Object Motion: Recognition, Learning, and Spatiotemporal Reasoning. *Journal of Robotics and Autonomous Systems* **8**(1–2): 65–91. Also in Walter Van de Velde, editor, *Towards Learning Robots*, The MIT Press, 1993.

Monmonier, M. (1991). *How to lie with maps*. University of Chicago Press, Chicago.

Montagnat, J., Delingette, H. & Ayache, N. (2001). A Review of Deformable Surfaces: Topology, Geometry and Deformation. *Image and Vision Computing* **19**(14): 1023–1040.

Montello, D. R. (1993). Scale and Multiple Psychologies of Space. In Frank, A. U. & Campari, I. (eds.) *Spatial information theory: European Conference COSIT '93*, 312–321, Marciana Marina, Elba Island, Italy, Lecture Notes in Computer Science 716, Springer Verlag. On-line at http://www.geog.ucsb.edu/~montello/scale.pdf.

Montello, D. R. (2001a). Scale in Geography. In Smelser, N. J. & Baltes, P. B. (eds.) *International Encyclopedia of the Social & Behavioral Sciences*, 13501–13504, Oxford: Pergamon Press (Amsterdam: Elsevier). On-line at http://www.geog.ucsb.edu/~montello/scale2.pdf.

Montello, D. R. (2001b). Spatial Cognition. In Smelser, N. J. and Baltes, P. B. (eds.) *International Encyclopedia of the Social & Behavioral Sciences*, 14771–14775, Oxford: Pergamon Press (Amsterdam: Elsevier). On-line at http://www.geog.ucsb.edu/~montello/spatcog.pdf.

Mukerjee, A. (1998). Neat Versus Scruffy: A Review of Computational Models for Spatial Expresssions, In Olivier, P. & Gapp, K.-P. (eds.) *Representation and Processsing of Spatial Expressions*, 1–35, Lawrence Erlbaum Associates, Mahwah, NJ.

Mukerjee, A., Gupta, K., Nautiyal, S., Singh, M. & Mishra, N. (2000). Conceptual Description of Visual Scenes from Linguistic Models. *Image and Vision Computing* **18**(2): 173–187.

Munkres, J. R. (1975). *Topology: A First Course*. Prentice-Hall, Englewood Cliffs, New Jersey.

Munkres, J. R. (1984). *Elements of Algebraic Topology*. Addison-Wesley, Redwood City, California and Reading, Massachusetts.

Murray, D. W. & Buxton, B. F. (1990). *Experiments in the Machine Interpretation of Visual Motion*. The MIT Press, Cambridge, MA.

Nagel, H.-H. (1988). From Image Sequences Towards Conceptual Descriptions. *Image and Vision Computing* **6**(2), 59–74.

Nagel, H.-H. (1994). A Vision of 'Vision and Language' Comprises Action: An Example From Road Traffic. *Artificial Intelligence Review* **8**(2–3): 189–214.

Neumann, B. (1989). Natural Language Descriptions of Time-varying Scenes. In Waltz, D. L. (ed.) *Semantic Structures: Advances in Natural Language Processing*, 167–206, Lawrence Erlbaum Associates, Hillsdale, NJ.

Neumann, B. & Novak, H.-J. (1983). Event Models for Recognition and Natural Language Description of Events in Real-world Image sequences. In *Proceedings of the Eighth IJCAI*, 724–726, Karlsrule, Germany.

Ng, J. & Gong, S. (2002a). Learning Intrinsic Video Content using Levenshtein Distance in Graph Partitioning. In Heyden, A., Sparr, G., Nielsen, M. & Johansen, P. (eds.) *Proceedings of the Seventh ECCV Conference*, Vol. IV. 670–684, Copenhagen, Denmark, Lecture Notes in Computer Science 2353, Springer Verlag.

Ng, J. & Gong, S. (2002b). On the Binding Mechanism of Synchronised Visual Events. In: *Proceedings of the IEEE Workshop on Motion and Video Computing*, 112–117, Orlando, Florida.

Novak, H.-J. & Neumann, B. (1986). Text Generation based on Visual Data: Descriptions of Traffic scenes. In Jorrand, P. & Sgurev, V. (eds.) *Artificial Intelligence II: Methodology, Systems, Applications — Proceedings of the Second International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA '86)*, 367–374, Varna, Bulgaria. North-Holland. Published 1987.

Oatley, K. G. (1977). Inference, Navigation, and Cognitive Maps. In Johnson-Laird, P. N. & Wason, P. C. (eds.) *Thinking: Readings in Cognitive Science*, 537–547, Cambridge University Press, Cambridge, England.

Olivier, P. & Gapp, K.-P. (eds.) (1998). *Representation and Processing of Spatial Expressions*. Lawrence Erlbaum Associates, Mahwah, NJ.

O'Regan, J. K. & Noë, A. (2001). A Sensorimotor Account of Vision and Visual Consciousness. *Behavioral and Brain Sciences* **24**(5): 939–1031.

O'Rourke, J. & Badler, N. I. (1980). Model-based Analysis of Human Motion Using Constraint Propagation. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **2**(6): 522–536.

Paoluzzi, A., Bernardini, F., Cattani, C. & Ferrucci, V. (1993). Dimension-independent Modeling with Simplicial Complexes. *ACM Transactions on Graphics* **12**(1): 56–102.

Paredaens, J. & Kuijpers, B. (1998). Data Models and Query Languages for Spatial Databases. *Data And Knowledge Engineering* **25**(1–2): 29–53.

Pasula, H., Russell, S., Ostland, M. & Ritov, Y. (1999). Tracking many Objects with many Sensors. In *Proceedings of the Sixteenth IJCAI Conference*, 1160–1167, Stockholm, Sweden.

Pentland, A. P. (1986). Perceptual Organisation and the Representation of Natural Form. *Artificial Intelligence* **28**: 293–331. Also in Fischler and Firschein (1987), 680–699.

Piaget, J. & Inhelder, B. (1956). *The Child's Conception of Space*. Routledge & Kegan Paul, London, England. An English translation of *La Representation de l'Espace chez l'Enfant*, Presses Universitaires de France: Paris, 1948.

Poincaré, H. (1958). *The Value of Science*. Dover Publications, Inc., New York. An English translation of *La Valeur de la Science*, Flammarion: Paris, 1905.

Poston, T. (1971). *Fuzzy Geometry*. Ph.D. thesis, Department of Mathematics, University of Warwick, UK. Microfilm.

Preparata, F. P. & Shamos, M. I. (1985). *Computational Geometry: An Introduction*, Springer-Verlag, Berlin. Expanded second printing, 1988.

Pylyshyn, Z. W. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. The MIT Press, Cambridge, MA.

Randell, D. A. & Cohn, T. (1989). Modelling Topological and Metrical Properties. In Brachman, R., Levesque, H. & Reiter, R. (eds.) *KR '89: Principles of Knowledge Representation and Reasoning*. Proceedings of the First Conference, Toronto, Ontario, Canada, pp. 357–368, Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Randell, D. A. & Cohn, T. (1992). Exploiting Lattices in a Theory of Space and Time. *Computers Mathematics Application* **23**(6–9): 459–476.

Rao, C., Yilmaz, A. & Shah, M. (2002). View-invariant Representation and Recognition of Actions. *International Journal of Computer Vision* **50**(2): 203–226.

Regazzoni, C. S., Ramesh, V. & Foresti, G. L. (2001). Scanning the Issue/Technology: Special Issue on Video Communications, Processing, and Understanding for Third Generation Surveillance Systems. *Proceedings of the IEEE* **89**(10): 1355–1367.

Roberts, F. S. (1973). Tolerance Geometry. *Notre Dame Journal of Formal Logic* **14**: 68–76.

Rosin, P. L. & Ellis, T. (1991). Detecting and Classifying Intruders in Image Sequences. In Mowforth, P. (ed.) *British Machine Vision Conference 1991*, 293–300, Glasgow, UK, Springer-Verlag, Berlin.

Rouke, C. P. & Sanderson, B. J. (1982). *Introduction to Piecewise-Linear Topology*. Springer-Verlag, Berlin.

Samet, H. (1984). The Quadtree and Related Hierarchical Data Structures. *Computing Surveys* **16**(2): 187–260.

Sapidis, N. & Perucchio, R. (1991). Delaunay Triangulation of Arbitrarily Shaped Planar Domains. *Computer Aided Geometric Design* **8**: 421–437.

Sato, T., Kanbara, M., Yokoya, N. & Takemura, H. (2002). Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-Baseline Stereo using a Hand-held Video. *International Journal of Computer Vision* **47**(1–3): 119–129.

Scaife, M. & Rogers Y. (1996). External Cognition: How do Graphical representations work?'. *International Journal of Human-Computer Studies* **45**(2): 185–213. Text on-line at http://www.cogs.susx.ac.uk as *Cognitive Science Research Paper 335*.

Schirra, J. R. & Stopp, E. (1993). ANTLIMA–A Listener Model with Mental Images. In *Proceedings of the Thirteenth IJCAI Conference*, pp. 175–180, Chambéry, France.

Schöne, H. (1984). *Spatial Orientation: The Spatial Control of Behaviour in Animals and Man*. Princeton University Press, Princeton, NJ. An English translation of *Orientierung im Raum* ... , published by Wissenschaftliche Verlagsgesellschaft: Stuttgart, 1980.

Schwartz, J. T. & Yap, C.-K. (eds.) (1987). *Algorithmic and Geometric Aspects of Robotics: Advances in Robotics 1*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Schweitzer, H. & Straach, J. (1995). Utilizing Moment Invariants and Gröbner Bases to Reason about Shapes. In: *Proceedings of the Fourteenth IJCAI Conference*, 908–914 Montréal, Canada.

Seitz, S. M. & Dyer, C. R. (1999). Photorealistic Scene Reconstruction by Voxel Coloring. *International Journal of Computer Vision* **35**(2): 151–173.

Sloan, S. (1987). A Fast Algorithm for Constructing Delaunay Triangulations in the Plane. *Advances in Enginnering Software* **9**(1): 34–55.

Sloman, A. (1975). Afterthoughts on Analogical Representation. In Schank, R. & Nash-Webber, B. (eds.) *Proceedings of Theoretical Issues in Natural Language Processing*, 164–168, MIT, Cambridge, MA, Association for Computational Linguistics. Also in Brachman and Levesque (1985), 431–439.

Sloman, A. (1985). Why We Need Many Knowledge Representation Formalisms. In Bramer, M. (ed.) *Research and Development in Expert Systems*, 163–183, Cambridge University Press, Cambridge, England. Proceedings of the Fourth Conference, University of Warwick, December, 1984. Also on-line at http://www.cs.bham.ac.uk/~axs.

Sonka, M., Hlavac, V. & Boyle, R. (1993). *Image Processing, Analysis and Machine Vision*. Chapman & Hall, London, England.

Spanier, E. H. (1966). *Algebraic Topology*. McGraw-Hill, New York. Second edition, Springer-Verlag: New York, 1989.

Stauffer, C. & Grimson, W. E. L. (2000). Learning Patterns of Activity Using Real-time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(8): 747–757.

Steels, L. (1988). Steps Towards Common Sense. In Kodratoff, Y. (ed.) *Proceedings of the Eighth ECAI Conference*. Munich, 49–54, London: Pitman Pub.

Steels, L. (1990). Exploiting Analogical Representations. *Journal of Robotics and Autonomous Systems* **6**(1–2): 71–88. Also in Pattie Maes, editor, *Designing Autonomous Agents*, The MIT Press, 1991.

Strat, T. M. & Fischler, M. A. (1991). Context-based Vision: Recognizing objects Using Information From Both 2-D and 3-D Imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(10): 1050–1065.

Sutton, R. S. (1990). Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Machine Learning: Proceedings of the Seventh International Conference*. Austin, Texas, 216–224, Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Talmy, L. (1983). How Language Structures Space. In Pick, H. L. & Acredolo, L. P. (eds.) *Spatial Orientation: Theory, Research and Application*. 225–282, Plenum Press, New York.

Tan, T. N., Sullivan, G. D. & Baker, K. D. (1998). Model-based Localisation and Recognition of Road Vehicles. *International Journal of Computer Vision* **27**(1): 5–25.

Tarabanis, K., Allen, P. & Tsai, R. (1995). A Survey of Sensor Planning in Computer Vision. *IEEE Journal of Robotics and Automation* **11**(1): 86–104.

Tarski, A. (1948). *A Decision Method for Elementary Algebra and Geometry*. Berkeley and Los Angeles: University of California. Second revised edition 1951.

Tarski, A. (1956). Foundations of the Geometry of Solids. In *Logic, Semantics, Metamathematics*, 24–29, Oxford University Press, Oxford, England. An address given to the First Polish Mathematical Congress, Lwów, 1927.

Thibadeau, R. (1986). Artificial perception of actions. *Cognitive Science* **10**(2): 117–149.

Thrun, S. (1998). Learning Metric-Topological Maps for Indoor Mobile Robot Navigation. *Artificial Intelligence* **99**(1): 21–71.

Toal, A. F. & Buxton, H. (1992). Spatio-temporal Reasoning Within a Traffic Surveillance System. In Sandini, G. (ed.) *Proceedings of the Second European Conference* *on Computer Vision*, 884–892, Genoa, Italy, Lecture Notes in Computer Science 588, Springer-Verlag.

Toffoli, T. & Margolus, N. (1987). *Cellular Automata Machines: A New Environment for Modeling*. The MIT Press, Cambridge, MA.

Tor, S. & Middleditch, A. (1984). Convex Decomposition of Simple Polygons. *ACM Transactions on Graphics* **3**(4): 244–265.

Toyama, K., Krumm, J. Brumitt, B. & Meyers, B. (1999). Wallflower: Principles and Practice of Background. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, 255–261, Kerkyra, Greece, IEEE Press.

Tsotsos, J. K. (2001). Motion Understanding: Task-Directed Attention and Representations that Link Perception With Action. *International Journal of Computer Vision* **45**(3): 265–280.

Tversky, B. & Lee P. U. (1998). How Space Structures Language. In Freksa, C., Habel, C. & Wender, K. F. (eds.) *Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pp. 157–175, Lecture Notes in Artificial Intelligence 1404, Springer-Verlag.

Udupa, J. K. (1983). Display of 3D Information in Discrete 3D Scenes Produced by Computerized Tomography. *Proceedings of the IEEE* **71**(3): 420–431.

Ullman, S. (1996). *High-Level Vision: Object Recognition and Visual Cognition*. The MIT Press, Cambridge, MA.

Wachsmuth, S., Socher, G., Brandt-Pook, H., Kummert, F. & Sagerer, G. (2000). Integration of Vision and Speech Understanding using Bayesian Networks. *Videre: Journal of Computer Vision Research* **1**(4): 62–83.

Weld, D. S. & de Kleer, J. (eds.) (1990). *Readings in Qualitative Reasoning about Physical Systems*. Morgan Kaufmann Publ. Inc., Palo Alto, CA.

Whitehead, J. H. C. (1949). Combinatorial Homotopy I. *Bulletin of the American Mathematical Society* **55**, 213–245. Also in *The Mathematical Works of J.H.C. Whitehead: Volume III on Homotopy Theory*, edited by I.M. James, Pergamon Press: Oxford, England, pp 85–177, 1962.

Worrall, A. D., Marslin, R. F., Sullivan, G. D. & Baker, K. B. (1991). Model-based Tracking. In Mowforth, P. (ed.) *British Machine Vision Conference 1991*. Glasgow, UK, 310–318, Springer-Verlag, Berlin.

Worrall, A. D., Sullivan, G. D. & Baker, K. B. (1994). A Simple, Intuitive Camera Calibration Tool for Natural Images. In *Proceedings of the Fifth British Machine Vision Conference*. BMVA Press, pp. 781–790.

Yap, C.-K. (1987). Algorithmic Motion Planning. In Schwartz, J. T. & Yap, C.-K. (eds.) *Algorithmic and Geometric Aspects of Robotics: Advances in Robotics 1*, 95–143, Lawrence Erlbaum Associates, Hillsdale, NJ.

Yeap, W. K. (1988). Towards a Computational Theory of Cognitive Maps. *Artificial Intelligence* **34**: 297–360.

Yip, K. & Zhao, F. (1996). Spatial Aggregation: Theory and Applications. *Journal of Artificial Intelligence Research* **5**: 1–26.

Zeeman, E. C. (1962). The Topology of the Brain and Visual Perception. In Fort, M. (ed.) *Topology of 3-Manifolds and related topics*. 240–256, Prentice-Hall, Englewood Cliffs, New Jersey.

Zhao, L. & Thorpe, C. (2000). Stereo and Neural Network-based Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems* **1**(3): 148–154.