Constraints, Descriptions and Domain Mappings in Computational Vision

A.K. Mackworth

Department of Computer Science, The University of British Columbia,
Vancouver, B.C. V6T 1W5, Canada

1. Introduction

The central paradox of computational vision is that given only one or more images of a scene the set of possible scenes depicted is underconstrained; however, our subjective experience is the opposite: the scene appears to be heavily overconstrained. Every aspect of our own visual experience offers mutually confirming evidence for the existence of a single, specific, non-ambiguous scene. This paradox can only be resolved by postulating that any perceiving system must supply organized knowledge of the scene domain, the imaging projection process, the radiometric and geometric aspects of lighting, the reflectance, transmission and refractance properties of scene materials and many other relevant physical regularities. This knowledge spans a spectrum from general *a priori* knowledge of all scenes in the domain assumed to be imaged - what can be and what can not - to the specific, contingent knowledge of the actual scene and imaging situation involved - what is and what is not.

2. Perception

All perception necessarily involves three domains: W, a set of possible worlds; I, a set of possible images of the world; and P, a set of possible projections of W into I.

A particular world, w, leaves a trace of itself, i, in I that is determined by the particular method of projection p. The triple of configurations in W, P and I is related by a relation of representation [1] R(w,p,i) as shown in Fig. 1.

Parenthetically, this formalism should not be narrowly interpreted as applying to, say, single perspective views of a three-dimensional world of opaque objects. The image domain could include motion primitives, aural or visual stereo pairs, haptic images or depth maps - whatever is directly sensed by the perceiver's transducers. The projection domain could include holographic processes, x-ray, PET, NMR or CAT imaging or even free hand sketching. The world domain includes three-dimensional worlds, maps, electric circuits, text and the like.

in Physical and Biological Processing of Images, (eds.) O.J. Braddick and A.C. Sleigh, Springer-Verlag, 1983, Berlin, pp 33-40.

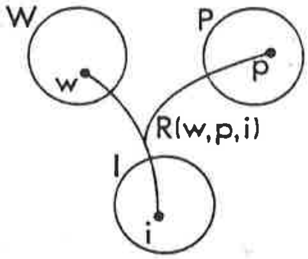

Fig. 1. The world, projection and image domains

The perceiver, minimally, has knowledge of i, I, W, P and R. Its task may be represented as finding a constructive proof for the imaging formula:

$$\exists w \exists p (w \varepsilon W) \wedge (p \varepsilon P) \wedge R(w,p,i) \quad . \qquad (1)$$

Any perceiver must be able to represent the potentially (and usually) infinite set of all solutions in a finite way. The representation must therefore be an intensional description of that set rather than an extensional list of its members.

In general the solutions will also be required to satisfy other constraints. Those constraints can be broadly categorised as a priori constraints on the nature of acceptable solutions in W or as arising from other images of the same w. Thus the representation must also allow specialization by additional constraints and intersection with other such representations. It might seem unnecessary to add the additional criterion that the representation include all and only those worlds that map to the given image i but many proposed representations have failed to satisfy this correctness criterion, in that they have excluded possible worlds for an image or included impossible worlds.

We assume that R(w,p,i) is restricted to the set of total many-to-one functional mappings from WxP onto I. This excludes for example, non-determinism from the imaging process (but not from the world). Thus, there is a mapping function,

$$f: WxP \text{ -> } I \text{ such that } R(w,p,i) \text{ iff } f(w,p) = i \ ,$$

f does not have a function as an inverse. However it is possible to partition the set WxP into equivalence classes such that (w_1,p_1) and (w_2,p_2) belong to the same equivalence class WP_i iff $f(w_1,p_1) = f(w_2,p_2) = i$. We have:

$$WP_i \subseteq WxP$$

$$WxP = \bigcup_{i \varepsilon I} WP_i$$

and $\quad i_1 \neq i_2 \text{ -> } WP_{i_1} \cap WP_{i_2} = \emptyset.$

We can project WP_i separately onto the set W as W_i and onto P as P_i where

$$W_i = \{w | (\exists p)(w,p) \epsilon WP_i\}$$

$$P_i = \{p | (\exists w)(w,p) \epsilon WP_i\} \quad ;$$

WP_i is a representation of the solution set of world-projection pairs for the imaging formula (1). Notice that if P is a singleton set $\{p\}$ (and the perceiver knows that) there is a one-to-one correspondence between members of WP_i and W_i. Since that assumption is often made, W_i is used to stand for the solution set of the simplified imaging formula:

$$(\exists w)(w \epsilon W) \wedge R(w,p,i) \quad .$$

The history of computational vision research has been a search for good representations of the equivalence class W_i which must be correct, incremental, finite and efficiently computable from i. Using those criteria the most satisfactory representations of W_i are sets of constraints, induced by i and p, on the set of possible worlds.

3. Vision

The general view of perception given above can be applied to vision in various ways. Both the world domain and the projection domain can be refined or factored into sub domains. A possible factoring starts with the "universe" domain, U, in which each universe is a set of objects possibly moving in space over time. An instant selected from the time domain, T, selects out from a universe a particular world in W. A particular lighting chosen from the illumination domain, M, of that world produces a lit world in L. A view from a particular direction (in V) of that lit world produces a scene in S while a projection of that scene produces an image in I. These domains and their relational mappings are shown in Fig. 2.

The ordering of the domains shown is somewhat arbitrary and could be changed for some applications. Indeed, another version of this scheme could eliminate some of the arbitrariness of ordering by allowing mapping relations among more than three domains. However the scheme presented has the same structure as the archetypal perception model introduced in Section 2. Each of the relations constrains elements in three domains and has the same characteristics as the original relation of representation, R. The convention shown in Fig. 3 indicates that the domain of relation R_n is AxBxC but there is also a functional mapping f_n: AxB -> C such that $R_n(a,b,c)$ iff $f_n(a,b) = c$. Indeed all of computer graphics is concerned with good representations for specific configurations in the domains shown and the mapping functions, f_n. As before, the mapping function induces an equivalence class partitioning of

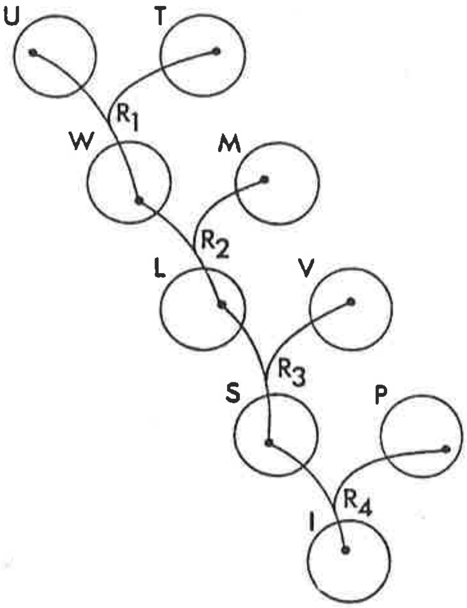

Fig. 2. The domains and relations for vision

AxB such that for any element $c \in C$ there is a class which we call $AB_c \subset AxB$ with $(a,b) \in AB_c$ iff $f_n(a,b) = c$. The fact that these equivalence classes are, in general, very large underlies tradeoffs such as surface photometry versus illumination or object shape versus projection technique that vary one factor while simultaneously varying another to produce identical images.

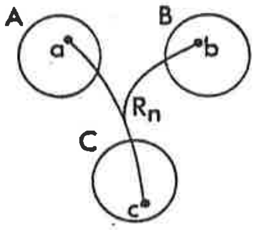

Fig. 3. The convention for showing R_n

Thus a given image, $i \in I$, induces an equivalence class partitioning $SP_i \subset SxP$. By projecting SP_i onto S and P respectively we have S_i and P_i. S_i in turn induces a partitioning of LxV, LV_i such that each pair (l,v) in LV_i maps into one member of S_i. LV_i can be similarly projected down into L_i and V_i and so forth.

Given the many-to-one nature of the mapping functions the size of the equivalence classes tends to increase with the level in the domain hierarchy unless other constraints on the classes are known a priori or available from other imagery known to relate to the same situation. Typical of the latter are motion, stereo and tomography, photometric stereo [2] and multispectral images which can each be illustrated as follows.

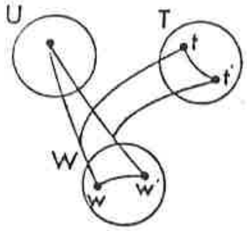

Fig. 4. Motion

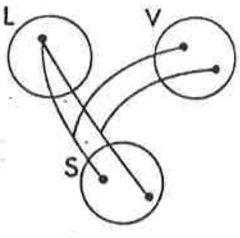

Fig. 5. Stereo

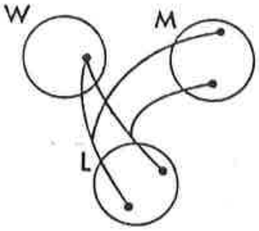

Fig. 6. Photometric stereo

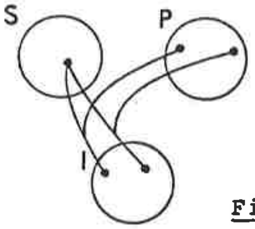

Fig. 7. Multi-spectral imaging

In multi-image situations that are known to arise from a single element of some higher domain one can exploit the extra constraints by intersecting the two (or more) equivalence classes that are induced in that higher domain: U for motion, L for stereo, W for photometric stereo, and S for multispectral imaging.

Most current theories of computational vision [3] distinguish between "viewer-centred" and "object-centred" representations. In this scheme the viewer-centred representations describe equivalence classes in S while object-centred representations describe equivalence classes in L and W.

The nature of each domain can be elaborated as much as necessary for any application. In a world of opaque objects, W, for example, could be composed of two domains, one for object geometry, the other for surface photometry, since they decouple into image geometry and radiometry.

In what follows, a brief sketch of some proposals for describing the equivalence classes in the scene domain, S_i, for orthographic line drawings of a world of opaque polyhedra will be given within this framework.

4. Some Scene Representations

When we say that a single image underconstrains the scene we mean simply that even given the knowledge that the projection is, say, orthographic the equivalence class S_i contains more than one member; in fact, it is obviously infinite and yet it is clearly a proper subset of S, the set of all possible scenes. To cope with this, various scene domain representations have been proposed. Here we can sketch five proposals and suggest weaknesses in each from the point of view of the criteria developed here.

The first representation uses Cartesian coordinate geometry. In a viewer-centred frame, let the image occupy the x-y coordinate plane and the projection be along the z axis. If the scene domain representation uses triples of reals as Cartesian coordinates, the scene representation fails the test of finiteness: the depth of the scene cannot be determined. To allow for that the second representation uses the equivalence class concept: represent all scenes which can be transformed into each other by a translation along the viewing direction by a single configuration. The group of translations along the z axis partitions S into a set of equivalence classes. Each equivalence class has an infinite number of members but it has a finite canonical representation as, say, the member of the class with a distinguished point having a z coordinate of zero. Unfortunately there is still an infinite number of equivalence classes that each map into a given image. We can now go on and look for another group of transformations which induce a further equivalence partitioning of our equivalence classes.

The third scene representation is the edge labelling approach [1,4]. The equivalence class descriptions are based on categorizing edges into four (or more) classes with respect to the dihedral angle the surfaces meet at and which view of the edge the viewer sees. Two scenes are then in the same equivalence class if they have the same visible edge and surface connectivitv and each edge has the same convex, concave or occluding label in each scene regardless of the actual dihedral angle between the two surfaces meeting at the edge. This equivalence class description is certainly finite; however, it is not correct. In general, not every member of the equivalence class produced can in fact map into the given image. This is always true but a convincing demonstration is given by "impossible objects" which are images that have no scene domain correspondent whereas the edge labelling algorithms return a non-empty equivalence class [4,5].

A fourth, more adequate, scene equivalence class representation is the gradient space representation of surface orientation [4,5]. This starts from the observation that equivalence classes induced by the translation group allow for one degree of freedom in the scene representation, but there are at least two more: the orientation of any surface in the scene can be arbitrarily specified. If there is a surface visibly intersecting that one another degree of freedom is introduced: the dihedral angle between them is unconstrained. However the locations of edge projections in the image supply a number of constraints in the scene that do constrain surface position and orientation in the scene thereby restricting the number of additional degrees of freedom in the scene description. A program POLY [5] based on this analysis, constructs equivalence class descriptions that are finite and eliminate many of the incorrect scene configurations allowed by edge labelling without, however, eliminating them all [6,7]. The descriptions are efficiently computable and are amenable to refinement or intersection with constraints available from a priori knowledge and multi-image constraints [8].

There is a general lesson in the usefulness of surface-based, viewer-centred intermediate scene representations. If a priori scene constraints such as "all surfaces are flat" or "all surfaces are smooth almost everywhere" are available. then it is necessary to find some configuration space in which such constraints are expressible. Minimally such a space must allow descriptions of the relevant objects (surfaces and edges, here) and constraints among the objects that are back-projected from the image description. Such a configuration space description is the requisite equivalence class description.

Another viewer — centred scene representation is the "intrinsic image" proposal [9]. Arrays of scalars are used to hold values at each image location for the various scene parameters that are confounded into image intensity such as surface illumination, albedo, orientation and depth. For our purposes here, the intrinsic image representation is not adequate. It does not satisfy the criteria of finiteness, correctness, refinability and efficient computability. It must be possible to construct finite equivalence class descriptions that describe the set of all scenes that could have produced the image and then refine that set as additional constraints become available. This process must use symbolic descriptions in the scene domain. Arrays of scalars can only represent one of the infinite number of members of the equivalence class. On the other hand, in an over-constrained imaging situation, where a particular member of several of the confounded domains is known and specified a priori, the intrinsic image representation is adequate.

5. Conclusion

The knowledge sources required to resolve the vision paradox can be constructed by understanding the confounding mapping processes discussed here. The needed equivalence class descriptions are best represented as sets of symbolic constraints on the allowable domain elements in a configuration space for the domain.

The general scheme is presented here to provide a framework for understanding theories of perception. It is not a theory of perception; however, any theory of perception must account for the fact that a perceiver does not perceive the world directly, pace Gibson [10]. The trace of the world available to the perceiver is a multiple confounding of configurations from many different domains. We see only the shadows dancing on the wall of Plato's cave.

6. References

1. CLOWES, M.B., On seeing things, Artificial Intelligence 2 (1971) 79-112.

2. WOODHAM, R.J., Analysing images of curved surfaces, Artificial Intelligence 17 (1981) 117-140.

3. MARR, D., Representing visual knowledge, AIM-415 (MIT, Cambridge, MA, 1977).

4. HUFFMAN, D.A. Impossible objects as nonsense sentences, in: B. Meltzer and D. Michie, Eds. Machine Intelligence 6 (Edinburgh University Press, Edinburgh, 1971) 295-323.

5. MACKWORTH, A.K. Interpreting pictures of polyhedral scenes, Artificial Intelligence 4 (1973) 121-137.

6. MACKWORTH, A.K. On the interpretation of drawings as three-dimensional scenes, D. Phil. Thesis (Sussex University, 1974).

7. DRAPER. S.W. The use of gradient and dual space in line-drawing interpretation Artificial Intelligence 17 (1981) 461-508.

8. MACKWORTH, A.K. Model-driven interpretation in intelligent vision systems, Perception 5 (1976) 349-370.

9. BARROW, H.G. and TENENBAUM, J.M., Recovering intrinsic scene characteristics from images, in: A. Hanson and E. Riseman, Eds., Computer Vision Systems (Academic Press, New York, 1978).

10. ULLMAN, S. Against direct perception, The Behavioural and Brain Sciences 3 (1980) 373-415.