Model Invocation for Three Dimensional Scene Understanding

Robert B. Fisher

Dept. of Artificial Intelligence, University of Edinburgh, Scotland

Abstract

Any gantral model-based vision system must somehow select a few serious candidates from its model base before applying model-directed processing. This is necessary for both efficiency and recognising 'similar* models (i.e. handling data errors, generic models and previously unseen objects). This paper shows how one can Integrate knowledge of object properties, structural and generic relations to create a network computation that performs model invocation. The paper demonstrates successful invocation in a scene containing a self and externally obscured PUMA robot.

1 In trod net ion

One important and difficult task for a general model based vision system is invoking the correct model. Because of the potentially huge number of possible objects, it is imperative that only a few serious candidates are selected for detailed consideration. Visual understanding must also include a pre-attentive element, because all models need be considered, yet active, direct comparison is computationally infeasible. Further, previously unseen objects, flexible objects seen in new configurations, incompletely visible objects (e.g. occlusion) and object variants (e.g. flaws, generics, new exemplars, etc.) require selecting models that are "close" to the data.

Model invocation is not just a visual problem (e.g. integrating cues while doing crossword pussies or invoking situation schemas), but here only the visual problem is considered. Invocation associates clues that suggest rather than verify. It may support the "seeing* of nonexistent, but highly plausible objects, as in surrealist art.

To date, little work has been done on sophisticated model invocation in the context of 3D vision. Using easily measured properties to select potential models fails in large model bases, because many objects share similar properties. Further, data errors, generic objects, object substructure and occlusion complicate indexing.

Arbib [I] proposed a schema-based invocation process with activation levels based on evidence, competition and cooperation from related activity. Marr [6) considered direct search in a model-base linked using specificity, adjunct and parent relations. Hinton and Lang [5] evaluated a connectionist model of invocation for 2D models, treating both model and data feature evidence identically. Feldman and Ballard [2] proposed a detailed computational model integrating evidence from spatially coincident property pairings.

This paper describes a solution that builds on these, embodying ideas on parallel networks, object description and representation in the context of 3D models and 3D visual information. The result is a plausibility calculation in a network structured according constraints defined by the structural, generic and context relationships.

Acknowledgements

Portions of this work were supported by a University of Edinburgh studentship. Thanks go to R. Beattie, J. Hallam, D. Hogg, J. Howe, M. Orr and many others.

2 Problem Context

These results are from the IMAGINE project [3] which investigated recognising 3D objects starting from 3D scene information. Earlier stages of processing include:

1. exploiting 3D feature continuity to overcome occlusion,
2. grouping individual surfaces to form primitive and depth aggregated surface clusters [4].
3. describing the significant scene features (curves, surfaces and volumes) by their 3D properties.

Model invocation happens at this point. After invocation, model-directed processes orient and verify the hypotheses.

Recognition starts from 2 1/2D sketch -like data segmented into surface patches of nearly uniform shape and separated by various shape or obscuring boundaries. As no well-developed processes produce this data yet, the program input is from computer augmented, hand-segmented test images. The paper illustrates the invocation process using a test image of a PUMA robot with its gripper obscured, using the surfaces shown in figure 1. The scene has flexibly connected rigid solids, has a variety of curved surfaces and the robot is both externally and self-obscured.

Object models are primarily structural, with features attached by reference frame transformations. The main primitives are the surface patch (SURFACE), characterised by consistent surface shape and polycurve boundary, and hierarchical groupings of surfaces (ASSEMBLY).

Four additional representations are added for invocation:

1. generic relationships between models,
2. major features of each assembly grouped according to viewpoint,
3. relevant properties with typical values for each structure, and
4. weighting factors modifying the importances of the properties and relationships.

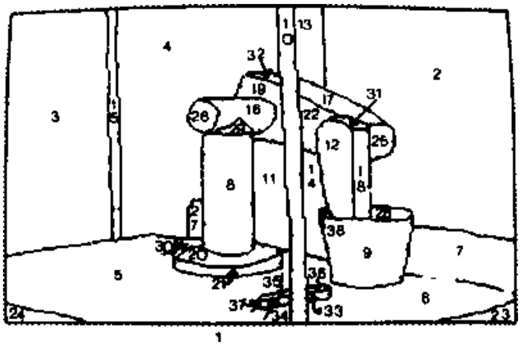

Figure 1: Test Scene Surface Regions

3 Theory: Evidence and Association

Model invocation is based on a plausibility value, in the range [-1,1], measuring how well an object model explains an image structure. Models are invoked when the plausibility is positive and sero is treated as a distinguished value. Plausibility is used because:

- some measure of similarity between objects is needed,
- it allows weak evidence support from associated hypotheses,
- it supports accumulating unrelated evidence types, and
- it degrades gracefully as data descriptions fail because of noise, occlusion or algorithmic limits.

The basic structural unit of invocation is the model instance • a given model in a given image context. Contexts are important because object features are usually connected or nearby. They define where image data can come from and what structures can provide supporting evidence. Here, the two types of contexts are the surface hypothesis and the surface cluster, which localise evidence for SURFACE and ASSEMBLY models respectively.

The inputs to invocation are:

- A set {C,} of image contexts.
- A set *{{d;,V,,Ct)}* of image descriptions of type *(d)* with value (v) for the features in these contexts.
- A database {(<,-,M^Mk$_f$w$_{tJ}$k)} of model-to-model (Af) associations of different types (t) with weights (w/).
- A database {(m{(d,y,/,,,tt$_{ti}$,w$_u$)})} of desired description constraints for each model, where *d* is the description type, [/,u] is the range of acceptable values and w is a weight.

The output of invocation is a set {{M,$_t$C$_{tt}$p$_{t)}$)} of plausibility measures for each model instance in each image context.

The plausibility of a hypothesis is a function of direct evidence from observed features and indirect evidence from associated hypotheses. For example, a toroidal shape is direct evidence for a bicycle wheel, whereas a nearby bicycle frame is indirect evidence.

Direct evidence is acquired from image data (e.g. 3D properties such as surface curvatures, axis orientations, etc.). Indirect evidence comes from associations, of which six types have been analysed here. The four distinguished types are supertype, subtype, supercomponent and subcomponent, the fifth is the competing identity relation and the sixth is the default association category. The nature of the influence depends on the type of the relationship. For example, an object necessarily requires most of its subcomponents to be present, whereas the reverse does not hold. Finally, a computation that integrates the seven evidence types to produce a single plausibility value is given.

When defining the different evidence computations below, a set of constraints and a satisfying function is given. Unfortunately, the constraints are never sufficient to implicate a unique function.

3.1 Direct Evidence

Direct evidence is calculated by comparing data properties (from within the appropriate image context) with model requirements.

Some constraints on this computation are:

- The contribution of a datum should be a function of its salience and the degree to which it meets its requirements.
- Each datum is considered only for the best requirement.
- Not all requirements need data (e.g. because of occlusion).
- Every datum must satisfy a constraint, if any of the appropriate type exist.

Based on these requirements, the following function evaluates a datum according to a requirement:

```
Let:
            n = nominal value (from model),
            r = nominal range (from model),
            w m importance weight (from model),
            d = data value
            e as evaluation
If:   | » - d | < r
            then:  e « w * (1 - 2 * ^=4)
            else: e *■ -u>
```

A description's evaluation is given by the requirement it fits best.

Because many models are likely to share some properties, any negative evidence should seriously reduce the plausibility, when integrating the individual evaluations. However, it should not cause immediate rejection, because the evidence may have arisen from (e.g.) partially obscured structures. Here, integration uses a weighted average that incorporates negative evidence doubly.

This algorithm defines a network fragment linking observed properties to model instances. The main processing elements are the property match evaluation, "max", "sum" and "weight" functions, with interconnections as defined by the above algorithm. The other algorithms below also have this property.

3.2 Supercomponent Associations

This association gives indirect evidence for a subcomponent, given the possible existence of an object of which it is a part. It is only suggestive, because the presence of the supercomponent implies that all subcomponents arc present (though not necessarily visible), but not that an image structure is any particular component.

Other constraints on the relation are:

- The more plausible the object, the more plausible its subcomponents.
- There is only one true superobject of any subcomponent.
- The object context must contain the subcomponent contexts.

One function meeting these constraints chooses the largest plausibility of any supercomponent in this or any larger context.

3.3 Subcomponent Associations

This association supports the presence of an object, given only the presence of its subcomponents (but not requiring any geometrical relationship). For a 3D object, only some of its subcomponents are typically visible from any particular viewpoint, implicating key feature groupings for integrating evidence.

Some constraints on this computation are:

- The more subcomponents present and the more plausible each subcomponent is, the more plausible the object is.
- Subcomponents are seen in viewpoint dependent groups (listing visible components according to salient viewpoint).
- The subcomponent context must lie within the object context.

The subcomponent computation occurs in three stages:

1. Find the most plausible candidate for each subcomponent in the given context and subcontexts.
2. Integrate the plausibilities of all subcomponents seen in each viewpoint grouping (weighted to express relative importance).
3. Pick the viewpoint grouping with the highest plausibility

3.4 Identity Inhibition

A structure seldom has more than one likely identity, unless the identities are generically related. Hence, an identity is inhibited by unrelated identities having high plausibilities In the same context. Inhibition also comes from the same identity in subcontexts, to force invocation to occur only in the smallest containing context.

Other constraints on the inhibition computation are:

- Only positive evidence for other identities inhibits.
- Inhibition varies with the plausibility of competing identities.

These constraints suggest the following computation:

- pick the largest plausibility of all generically un-related hypotheses in the same context and all hypotheses of ths same type in sub-contexts. Call this *P*.
- if *P* > 0, then the inhibitory plausibility is *-P;* otherwise no inhibition is applied.

3.6 Evidence Integration

The seven evidence types are integrated to give the plausibility value for the model hypothesis. Some constraints on this computation are:

- Directly related evidence (direct, subcomponent and subtype) should have greater weight.
- Other indirect evidence should be somewhat incremental.
- Only types with evidence are used.
- If there is no direct, subtype or subcomponent evidence, then evidence integration produces no result.
- Direct and subcomponent evidence are equivalent and the weaker of the two should be taken.
- Strong supercomponent or association evidence supports.
- The plausibility of a type must be at least that of the subtype and at most that of the supertype.
- If other identities are competing, they inhibit the plausibility.

One function meeting these constraints is:

Let:

```
$e_{dir}, e_{subt}, e_{supt}, e_{subc}, e_{supc}, e_{ass}, e_{inh}$
be the seven evidence values.
```

Then:

```
$v_1 = min(e_{dir}, e_{subc})$
if $e_{supc} > 0$
        then $v_2 = v_1 + c_{supc} * e_{supc}$        $(c_{supc} = 0.1)$
        else $v_2 = v_1$
if $e_{ass} > 0$
        then $v_3 = v_2 + c_{ass} * e_{ass}$        $(c_{ass} = 0.1)$
        else $v_3 = v_2$
if $e_{inh} > 0$
        then $v_4 = v_3 + c_{inh} * e_{inh}$        $(c_{inh} = 0.25)$
        else $v_4 = v_3$
Finally, the integrated plausibility value $p$ is:
        $p = min(max(v_4, e_{subt}, -1.0), e_{supt}, 1.0)$
```

Table 1: Invoked Hypotheses

MODEL	REGIONS	PLAUS	ST	NT
shoulder body	16,26	0.45	E	
trash can	9	0.34	E	
shoulder	16,26,29	0.28	E	
shoulder small panel	29	0.26	E	
shoulder small panel	27	0.24	I	3
lower arm	12,18,31	0.23	E	
shoulder small panel	20,21,27,30	0.20	I	3
body	9,12,17,18,19,22 25,26,31,32,36	0.18	I	1
body	9,12,18,28,31,36	0.18	I	1
body	8,9,12,16,17,18 19,22,25,26,28 29,31,32,36	0.17	L	
body	9,28,38	0.17	I	1
link	8,9,12,16,17,18 19,22,25,26,28 29,31,32,36	0.08	E	
link	8,16,17,19,20 21,22,25,26,27 29,30,32	0.06	I	4
upper and lower arm assembly	9,12,17,18,19,22 25,26,31,32,36	0.05	E	
link	8,9,12,16,17,18 19,22,25,26,28 29,31,32,36	0.04	E	
upper arm	17,19,22,25,32	0.03	E	
lower arm	17,19,22,25,32	0.00	I	2

STATUS
E - invocation in exact context
L - invocation in larger context than necessary
I - invalid invocation

NOTES
1 - because trashcan outer surface very similar
2 - similarity with upper arm model
3 - ASSEMBLY with single surface has poor discrimination
4 - not large enough context to contain all components

4 Evaluation

Ideally, invocation should select all correct models in only the correct context and the only false invocations are those "similar" to the true ones. Further the computation should converge. As the network structure is scene dependent and non-linear, only minor mathematical results have been found so far, concerning:

- relative ranking between the same model in different contexts,
- correct direct, subcomponent and generic evidence always implies invocation, and
- the number of direct evidence properties to use.

So, a performance demonstration is presented, using a network created from the test scene, showing that, here, invocation is effective.

Table 1 lists all invoked ASSEMBLY models and the associated image regions, ranked by plausibility. (Invoked surface models are not listed.) These were computed in a network containing 252 model-data pairing nodes for 14 ASSEMBLY models in 18 surface clusters, and 425 nodes for 25 SURFACE models in 17 data surfaces.

Invocation was selective with 17 ASSEMBLY invocations of a possible 252, where 10 of the 17 invocations were correct and 0 were in the smallest appropriate context. All appropriate invocations occurred. Of the incorrect, only 3 were unjustified (notes 2 and 3 in table 1). From the SURFACE model invocations (not shown), 24 invocations were made out of 425 possible. Of these, 10 were correct, 10 were justifiably incorrect because of similarity and 4 were inappropriate invocations. Clearly, invocation worked well here.

The chief causes for improper invocation were:

- not large enough context to contain all subcomponents while having structures not contained in the successful context, and
- superficial similarity between features.

5 Discussion

The invocation process discussed here integrates the major evidence types and eliminates the need for exhaustive model-directed testing while still maintaining access to all models. By basing invocation on propagated plausibility values, the data to model comparison computation has been regularised, and is amenable to parallelism.

To summarise, this paper has:

- defined a formal basis for visual model invocation.
- proposed constraints on the effect of different evidence types.
- proposed surfaces and surface clusters as the contexts in which to consider invocation.
- demonstrated a successful implementation of the theory in a simplified 3D context.

6 Bibliography

1. Arbib, M. A., *Local Organising Processes and Motion Schemas in Visual Perception,* in Hayes, et.al., Machine Intelligence 9, pp287-298, 1079.
2. Feldman, J. A., Ballard, D. H., *Computing With Connections,* in Human and Machine Vision, Beck, Hope and Rosenfeld (eds), Academic Press, pp 107-155, 1983.
3. Fisher, R. B., *From Surfaces to Objects: Recognising Objects Using Surface Information And Object Models,* PhD Thesis, Dept. of Artificial Intelligence, Univ. of Edinburgh, 1986.
4. Fisher, R. B., *Identity Independent Object Segmentation in 2 1/2D Sketch Data,* Proc. 1986 European Conference on Artificial Intelligence, July 1986.
5. Hinton, G. E., Lang, K. J., *Shape Recognition and Illusory Connections,* IJCAI 9, pp252-259, 1985.
6. Marr, D., Vision, W.H. Freeman and Co. (pub), 1982.