To appear June 1992

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS

A Lattice Framework for Integrating Vision Modules

Allan Jepson
Whitman Richards

Abstract: Because of the successes in understanding information processing by individual modules such as stereopsis, motion, texture and color, research in computational vision is now turning to studies of how information provided by these modules may be integrated or "assimilated". We propose a framework for assimilation based upon a partial ordering of constraints implicit in all active modules. Such constraints, for example the rigidity constraint for motion, although often robust are also fallible, and hence are more properly regarded as premises. Such premises are used to construct a preference ordering for (classes of) interpretations of the image. Interpretations associated with maximal states in the ordering are taken as the assimilated interpretation of the modules. This approach stresses the need to use world knowledge to reason about the plausibility and consistency of interpretations of the image data.

Allan Jepson was supported by NSERC Canada. Whitman Richards was supported by AFOSR 89-504.

Allan Jepson is with the Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Ontario M5S 1A7. Whitman Richards is with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 79 Amherst St., Cambridge, Massachusetts 02139.

1.0 Introduction

Research in both machine and biological vision has typically proceeded with a "divide and conquer" strategy. Hence, strong emphasis has been placed upon understanding the components of the visual information process, such as stereopsis, motion, texture, and color. This approach naturally leads to a modular view of vision [12], [40], [41]. Because of the advances in understanding such modules both at a theoretical and applied level, recently attention has turned to integrating or "assimilating" their outputs [3], [20], [33], [34], [51], [52] [56]. Such studies of how modular information may be combined, especially in the presence of conflicting or ambiguous data, fall into three classes: 1) The removal of ambiguities by one module constraining the other (e.g. stereo and structure-from-motion [54], [67]; stereo and shading [26]; shading and motion [2] [3]; shading and texture [48]); 2) Assigning a priority to each module, enabling the one with the higher priority to veto the conclusions of the other having lower priority (e.g. stereo and shading [18]), and 3) Taking a vote, such as the Hough transform [8], [9], or weighted probabilities [22], [42], [46], [58], [61]. Here we propose still a fourth approach, based upon a partial ordering of constraints implicit in the modules.

Our basic idea is that the evidence extracted from images for the structure of the scene can be interpreted using fallible premises [35]. For example, if an image feature is observed which is considered to be evidence for the rigid 3D motion of an object, then we may wish to use the premise that the image feature does indeed depict the particular 3D motion. This is necessarily a fallible premise, since any measurable image feature could have arisen by a coincidence. Given a pair of situations in which such a premise is held in one and rejected in the other, all other things being equal, then we prefer interpretations in which the premise is held and the evidence is accepted. There will typically be many premises of this character. These premises are used to construct a preference ordering for (classes of) interpretations of the image. Of course, conflicts will arise between various premises; for example, in situations in which it is inconsistent to keep all of one's premises. Such inconsistent sets of premises are deleted from our ordering. As a consequence, the partial order may not have a global maximum. Finding a local maximum in the ordering of sets of fallible premises can be viewed as the integration or "assimilation" of information offered by different vision modules. Here we introduce a framework for such an approach, leaving aside algorithmic considerations.

2.0 Some Definitions

What is a module? Our intuitive notion is a chunk of information processing capability which uses constraints to organize aspects of the sense data, out-putting an interpretation of the data as a world structure or event. (See Reiter & Mackworth [53] for a formal definition of an interpretation.) Thus, the structure-from-motion module takes in "N" corresponding points in "M" frames, postulates rigidity or fixed-axis motion, etc., and outputs a possible 3D structure and motion of this structure in the world. By choosing different constraints, or premises as we shall call them here, one can change the plausible interpretations of the data. A module is thus a triple (M, P, O) consisting of a set, M, of measurement operations applicable to the (sensory) data, a set, P, of premises suitable for the interpretation of these measurements, and a partial order relation O for collections of these premises. Given its input data[1], the principal output of a module is an interpretation I which is maximally preferred according to the partial order O. This interpretation is not necessarily unique. (Consider the reflection ambiguity in Ullman's structure-from-motion module, or the two states of the simple Necker cube.)

What information is passed on by a module? Contrary to one's first thought, we do not propose that a module simply pass on an interpretation of the sense data. It is here, then, that our proposal differs radically from previous schemes for integrating modular information by use of vetoes, priorities, probabilities, etc. applied to the modular interpretations. A module cannot solely output its conclusions. Instead, we must realize that such conclusions are inductive inferences based upon premises (i.e. constraints) which are fallible. Consequently, when two or more modules are to participate in arriving at an interpretation of the sense data S, they must be able to resolve conflicts in their interpretations at the level of the premises used to arrive at any conflicting interpretations [13]. The sense data are inviolate; the premises are not. Hence information about premises must be passed between modules. This is a key to understanding how our scheme for modular integration works.

3.0 The Proposal

Everyone is familiar with 2D structure-from-motion displays which yield vivid 3D perceptions. Our TVs provide the most common examples. The striking fact about such displays is that we see 3D structure in the face of conflicting information from stereopsis. The disparity provided by our binocular system must certainly output zero, indicating a 2D planar display. Yet this is not what we see. Has the structure-from-motion module simply vetoed the stereopsis channel, perhaps with the assistance of a shape or form module? If so, then what are the rules that underlie when a veto should be exercised or not? Here, we propose a scheme which superficially might look like a veto process, but more properly is an inductive inference across the modules. An example will help clarify our proposal.

Consider only the most minimal structure-from-motion display consisting of moving points. For the structure-from-motion (SFM) module to output a 3D interpretation, it might use Ullman's rigidity constraint [64]. Thus, we know if the display consists of at least four points in motion, then at least three views will provide a unique interpretation of these points as a rigid 3D structure, provided the 3D points are not co-planar. Hence, the assignment of a 3D structure to the image points is the interpretation offered by the SFM module, given its rigidity premise. But, of course, this premise is fallible.

Consider now the stereo module as it "views" the flat TV screen. It must output a planar, non-rigid interpretation for the 3D structure because the point disparities indeed are all co-planar. The simplified underlying premise is that disparities indicate relative or ordinal distances. However this premise is also fallible, as when we view a very distant object on the horizon.

To summarize, we have the following for a "frontal view" of the TV display:

Structure-from-Motion module:

R: ***Rigidity Premise***: Image motion consistent with a rigid 3D object depicts that rigid object in motion.

Stereo Module:

D: *Coplanarity Premise*: A constant gradient of disparity depicts a planar surface.

——— *Figure 1 about here* ———

We can begin to assimilate these results by constructing a lattice, where the nodes of the lattice are the possible combinations of the premises R (rigid), and D (disparity) and their faults $\overline{R}$ and $\overline{D}$, indicating that the premise is "given up". Figure 1 shows such a lattice. The topmost node has the interpretation of the points as rigid and coplanar. However such an interpretation is inconsistent – given the data and the premises there is no world structure consistent with both. Hence this node is deleted by cross-hatching. At the opposite extreme is the node $\overline{R}$ and $\overline{D}$ where both the rigid and coplanar premises are given up. This node is not cross-hatched because there are indeed many possible 3D world structures that could generate such point image motion. For example, we could move the points arbitrarily in and out along the line of sight and place the structure at a considerable distance so the disparities are still zero.

Remaining are the two nodes $\overline{R}D$ (non-rigid, coplanar) and $R\overline{D}$ (rigid, non-coplanar). These nodes are respectively labelled "Stereo" and "Motion" because they correspond to the conclusions reached separately by these two modules. At the heart of our proposal is the simple observation, that given everything else being equal, we should prefer interpretations for which our premises are satisfied over interpretations for which they are violated.f[2] Indeed, for the two premises R and D, the image measurements used in each can be considered evidence for the corresponding scene structure. Therefore, the premises induce an ordering of the interpretations, as is indicated by the arrows in Figure 1. In particular, interpretations consistent with either $\overline{R}D$ or $R\overline{D}$ will be preferred over $\overline{RD}$. Given our two premises, these are the so-called "maximal" states. So how do we decide which of these to pick as our perception? Is it possible to achieve a unique maximal node within the existing framework? Can we simply add more constraints or premises and thereby force a unique maximum? Surprisingly, as we illustrate below, adding extra premises will not decrease the number of local maxima.

——— *Figure 2 about here* ———

Let "X" be a premise used to augment a lattice such as we have done in Figure 2. For example, "X" might be the premise that objects rotate about fixed axes (even when non-rigid); or "X" might be the premise that the object possesses some kind of symmetry. (For symmetry presumably we would have to query a "shape" or "form" module; whereas for fixed axis motion, perhaps this premise also resides in the SFM module along with rigidity.) Now regardless of where these new premises reside, any node in the lattice with both the R and D premises true must be deleted as before. So the issue is whether the new premises can annihilate enough of the lattice of Figure 2 so the resulting lattice of consistent nodes has a unique maxima. Specifically, if the Motion node $RX\overline{D}$

is to become maximal, then both $\overline{R}XD$ and $\overline{R}\overline{X}D$ must become inconsistent for the added premise "X".

——— *Figure 3 about here* ———

Consider now the case of fixed axis motion, with a fixed-axis premise:

F: *Fixed Axis* : Image motion consistent with a 3D fixed axis motion depicts 3D fixed axis motion.

Clearly the previous Motion interpretation (resident in the $RF\overline{D}$ node) remains intact, but the original Stereo interpretation associated with the $\overline{R}D$ node now becomes $\overline{R}FD$ which is inconsistent (for images generated from a generic view). As shown in Figure 3, given D both R and F must be faulted, whereas for the Motion module, only one premise D must be given up as before. However, it is important to note that the $RF\overline{D}$ node is *not* a unique maximal node in our lattice based on the independent preferences for maintaining R, D, and F. Furthermore, given any other premise, X, to be used instead of F, the resulting lattice will also contain (at least) two local maxima.[3] If a unique local maximum is desired then we must consider something beyond adding another premise.

——— *Figure 4 about here* ———

There are several approaches for enhancing the lattice in Figure 3 in order to obtain a unique maximum. We illustrate three of these in Figure 4. First we consider a voting scheme [8], [9]. In Figure 4A (top panel) we have drawn a dashed arrow from the 2-fault Stereo node to the 1-fault Motion node because the latter contains more votes for our premises. (A similar result would occur if evidence for rigid fixed axis motion always vetoed the disparity information.) But what grounds do we have for making this maneuver? Implicit is the assumption that the evidence supporting the rigid-fixed axis motion interpretation (RF) is *always* more likely to be true than the evidence supporting stereo disparity (D). Clearly we're making a bet based on probabilities.

However, if we resort to probabilities to force a unique maximal node, then it's possible, in the absence of any hard constraints about the world, that evidence for stereo disparity has a higher probablity of being valid than rigidity. In this case, if the ordering is based on probabilities, the lattice will be as illustrated in Figure 4B (middle panel), with the stereo node maximal. So probabilities can result in either lattice 4A or 4B, depending upon one's assumptions about the distribution of events in the world. What we are striving for here is a partial ordering that is robust under a wide range of different probabilities of various world events.

Consider again just what is known when we view a TV image of a rotating 3D rigid object. Our stereo module asserts that the display is in a plane (so D is satisfied). On the other hand the motion of the points in this plane provide evidence for the rigid fixed-axes motion hypothesis (so R and F are satisfied). In the presence of such conflicting evidence, why not ask whether there is a class of world structures that offers an explanation for both? In this spirit, we introduce the concept of a picture and its associated premise as follows:

> **P**: *Picture Premise*: Given stereo disparity consistent with a planar surface (not at infinity), and given other evidence for 3D (non-planar) structure, then the image depicts a picture of the 3D structure.

The effect of adding both the concept of a picture and its premise P is illustrated in Figure 4C (bottom panel). Now, when together with P, the premises RFD become consistent and $(PRFD)$ will be a unique maximal node. For this node the stereo data are interpreted as providing the depth information for the picture; whereas the premises R and F provide constraints on how the picture is interpreted. Note that the critical step was to recognize that the inconsistency of the other forms of evidence for 3D structure was the evidence for a picture. If this example of a TV screen may seem a little artificial, we need only consider the interpretation of stereo disparity in scenes containing reflecting objects (such as calm water, metal, or pottery) to realize that similar concepts might also be required to understand many images in the natural world.

4.0 Representation by Parts

In our previous example, seemingly conflicting conclusions reached by the motion and stereo modules were assimilated by considering how the active premises about world structures (i.e. R and D) constrained how the sense data was to be interpreted by these modules. It was seen that adding the concept "picture", and its associated premise P, we could resolve the conflict and reach a unique, global maximum in a fault lattice constructed from premises RDP. Clearly, however, in the complexity of the natural world, there will be many cases where our repertoire of concepts will either be too vast or will be insufficient under such conditions. It is unreasonable to require that we conduct an exhaustive search over all our models and concepts, with their affiliated premises, when a unique global situation is not guaranteed. What then is a strategy for assimilation?

One obvious strategy is that a working ordering be given to the premises: those premises which are typically found to be most useful are invoked first. For example, almost all schemes for object representation are part-based [5], [7], [11], [14], [30], [40], [43], [49], [62]. However, part-based representations alone are not sufficient because some objects and configurations, such as crumpled paper or smoke, do not lend themselves easily to part-based descriptions. Nevertheless, the use of parts is a very natural and efficient decomposition which greatly simplifies how image features should be grouped. Here, we explore the consequences of assimilation driven by such an approach. This is done in the context of a second example, the Ames Trapezoid Window.

4.1 The Ames Trapezoid Window

In 1951, Ames [4] presented a paradoxical demonstration which exhibited a striking failure of the rigidity constraint for the structure-from-motion theorem (see Hochberg [32], for other, similar failures in veridical perception). A rod is placed diagonally in 3D through a trapezoid window, and attached rigidly to the crosspieces in the window. The four corners of the trapezoid (plus its internal crossbar structure) together with the two end points of the rod thus satisfy Ullman's rigidity constraint (see Figure 5). Yet when the window is rotated and viewed in 2D on a video screen the structure appears non-rigid. The rod seems to follow its own rotating motion separate from the perceived rotation of the window, which is pendular with the short side always to the rear. Although there is indeed a unique 3D interpretation of this configuration as a rigid structure, we do not perceive it. (We have created a similar illusion by attaching a stick "handle" to one face of a wire-frame cube. In 2D viewing, the "rigid" handle appears to flop around as the cube rotates.)

Putting aside the problems of stereo views of 2D structure-from-motion displays (whose resolution follows the earlier example), we direct our attention to the role a parts-based description of the display will play in the construction of a fault-lattice of premises associated with "Form" and "Motion" modules. For this example, we will see that the lattice built from a parts-based description will have two maximal nodes, one containing the (correct) rigid interpretation, the other having a non-rigid interpretation. Indeed, more generally, we show that a part-based decomposition of an image region containing any number of rigid parts will always produce a maximal node with a rigid interpretation. Thus, a representation by parts is not destructive in that a node containing the "correct" rigid interpretation is not annihilated.

4.2 The Structure-from-Motion Module

Although the "rigidity" premise analyzed by Ullman [65] is a very popular premise for recovering the 3D structure of a collection of moving feature points, there are several other premises that have been found equally useful [1]. Among these are those for articulated motions [29], [57]; fixed axis motions [16], [28] and more simply, planar motions [29]. However all these other proposals have one common element: the parts comprising the structure are generally rigid. Hence here for simplicity we take as an axiom that any part-decomposition must be such that the part is a rigid body.

By implication, a part decomposition of an object suggests that various parts are attached to one another. We note that for the Ames display the points of attachments are stable under motion, just like they are for articulated motions. Also, as is the case for the common Ames display, we assume that the fixed axis rotation is nearly parallel to the plane of the display.[4]

Thus, for simplicity, we have taken as hard constraints the following two axioms which in a full treatment would more properly be treated as fallible premises:

1. Parts are rigid.
2. Fixed axis rotation.

Finally, we associate with the motion module the following fallible premise which is a specialization of that used earlier:

> **R**: *Rigidity Premise*: If the image motion of two parts attached to each other is consistent with a rigid 3D motion, then those parts depict that rigid configuration.

So in the case of Ames trapezoid plus bar, which indeed satisfies the rigidity constraint, R may be restated as "the bar is rigidly attached to the trapezoid window". In support of this premise we note here that if the SFM module is isolated from the form module by eliminating all lines and reducing the configuration to a collection of moving points (Figure 5, top), then the common percept is indeed of rigid 360 deg motion.

——— *Figure 5 about here* ———

4.3 The "Form" Module

The idea behind our choice of the premises for this module is that a single, static view of a shape can induce the appearance of a slant, as has been well documented [23], [59]. For example, the lower two panels of Figure 5 show the Ames window plus bar broken down into two "objects", or "parts", namely the window itself (left) and the extra bar together with the horizontal crossbar of the window (right). (We assume the grouping of line segments comprising the window is due to the non-accidental alignments of the parallel and symmetric lines, as suggested by Binford [15] and Lowe [37]). Each of these objects alone can be assigned a local coordinate frame, with a slant and tilt specifying the orientation of the frame with respect to the viewer. For the bar plus horizontal crossbar of the window (right) the impression, as shown in Figure 5, is of a planar surface viewed from above. We assume that the slant and tilt of this plane is calculated along the lines proposed by Stevens [60]. Call this "Form" premise that specifies the partial coordinate frame associated with bar as premise B. Similarly, the trapezoid window itself can specify a (different) coordinate frame, again using Stevens' rules, or perhaps some other consideration such as that proposed by Brady & Yuille [17] or Kanade [36]. Call this premise W. Because the window has one short vertical side, this will be taken as located behind the larger vertical side, causing the slant of the window always to be positive. Our "form" premises are then the following:

W: The trapezoid window is slanted with the short side away from the observer.

B: The bar is attached to the horizontal crosspiece of the window, with the pair seen from above.

Again for simplification, we will take as axiomatic, bolstered by the symmetries and alignments of the window, that the window is planar. In sum, the additional simplifying axioms are:

3. The window and bar are parts.
4. The window is planar.

In support of the premises W and B associated with the form module, consider the lower panels of Figure 5. If only the trapezoid window is depicted (Figure 5, left), then it is hard not to see this configuration as a slanted plane. Similarly, when the bar is isolated as a part attached to the cross-piece of the window, the common impression is of two lines lying on a near horizontal plane, viewed from above (Figure 5, right).

4.4 Fault Lattice for RWB

To interpret the Ames Window display when two (or more) modules are simultaneously active and delivering their conflicting conclusions, we proceed as before and first construct a lattice of the three premises R, W, and B and their faults. This lattice is shown in Figure 6. Note that now the modular associations of the premises become unimportant.

——— *Figure 6 about here* ———

First consider the top-most node of the lattice RWB where no premise is given up or faulted. Clearly this is not possible for the Ames' display, for the coordinate frames and motions by W and B are different, and hence the rigidity premise R can not hold. The same arguments apply to the premise pairs RB and RW, because the structure can not be both articulated and rigid at the same time. Thus any node containing RB or RW unfaulted can be deleted as inconsistent. These exclusions leave only node $\overline{R}WB$ as valid at the level in the lattice where nodes have only one fault. Thus, this leaves us with node $\overline{R}WB$ as the single maximal node. This node contains an interpretation (1) of the display which is of a non-rigid configuration ($\overline{R}$), with the coordinate frame and motion of the bar different from the coordinate frame and motion of the window. This "causes" the bar-frame to be seen as articulated with respect to the window frame, which undergoes pendular oscillation with its short side to the rear. This interpretation (1) is the most commonly observed percept.

Consider now the inconsistent node $RW\overline{B}$. This node has two children, each with two faults. One, $R\overline{WB}$, is a consistent node with a rigid interpretation (2). This is allowed here because the bar and window coordinate frames are faulted, along with their viewpoint restrictions. Note that this node is also a maximal node because there is no path to a higher consistent node.

There are two more consistent nodes at the two-fault level in the lattice. One, $\overline{R}W\overline{B}$ is the other child of the inconsistent node $RW\overline{B}$. This node contains a non-rigid interpretation (3) of the bar rotating 360 deg (causing a violation of "viewed from above") with the window still undergoing pendular oscillation. This interpretation is less favored over (1) which is affiliated with the one-fault node $\overline{R}WB$. Our explanation for this preference is simply that $\overline{R}W\overline{B}$ is a child of a consistent higher node with fewer faults. Similarly, the two-fault node $\overline{RW}B$ is also a child of $\overline{R}WB$ and contains still another, less seen interpretation (4) of the window undergoing 360 deg rotation, but with the bar articulated with pendular motion.

Finally, the lowest, three-fault node also contains consistent non-rigid interpretations, one of which is (5) with the bar lying in the plane of the window, and the window rotating the full 360 deg. This interpretation is very difficult for most observers to see. According to our theory, this is not suprising because paths lead upward to several other nodes that satisfy more of our premises.

Thus, the top of the RWB fault lattice contains two locally maximal nodes. One, node $\overline{R}WB$, has only one fault and holds the commonly perceived non-rigid interpretation of a stick wobbling in the window. The other locally maximal node, $R\overline{WB}$, has two faulted premises, and is the interpretation of the entire configuration as rigid. However, the higher maximal node with the fewer faults contains an incorrect interpretation. Has assimilation then failed?

4.5 Competence Versus Performance

We see the primary job of an assimilator as explaining the sense data, or in this case the conclusions reached by its modular inputs, in terms of world events [25], [55]. In the case of the Ames illusion, there were two explanations found, each associated with different maximal nodes. Indeed, with our current set of premises, if we invoked N rigid part decompositions of the image data in a region with n

pairwise attachments, we can expect up to 2^n different local maxima, or "explanations". Among these will be one for a non-articulated, rigid configuration if indeed that is what is present in the world, as in our example. Such an interpretaion is guaranteed in this case because all pairs RX, where X is a premise for any one of the possible rigid but articulated parts, will be inconsistent. Hence $R\overline{XY}\ldots$ must be a maximal node for any completely rigid 3D configuration. The problem the assimilator has is to choose among the maximal nodes. Clearly more information may be needed to decide "correctly". This information might include premises or measurements from other modules as well as the ordering relations and premises of the current assimilator. Thus the assimilator itself is also a triple $[M, P, O]$ where M, P, and O at least contain the union of the corresponding sets for the modules being assimilated. In this view, an assimilator has the same structure as a module as defined in Section 2.0 [13]. However, we do not wish to imply that assimilators or modules at the highest levels have direct access to all the premises at lower levels. The part-based decompostion used earlier would be one example where the output of an earlier (part-selection) module whose premises were taken to be not directly accessible.

——— *Figure 7 about here* ———

As mentioned above, when faced with multiple maxima, we have two basic options: (1) more data together with associated premises can be sought, or (2) a stronger decision rule or ordering may be chosen. An example of this second option would be a voting rule discussed earlier. There is a third alternative, however, more in the spirit of trying to explain the data.

Consider again the implications of a parts-based strategy for image interpretation. In the Ames example, our parts were taken as the window and the bar. If these groupings of the image data into "objects" or "object-parts" is obvious, then why not simply explore the consequences of these groupings in light of the data? Clearly this strategy generally works and will avoid an enormous potential search over concepts, premises, and other possible part-decompositions [63]. The fault lattice will then have the structure illustrated in Figure 7 (left), where W and B are the "window" and "bar" premises used earlier. In this lattice, each node has a consistent interpretation, indicated by the numbers, which are the same as in Figure 6. Because all paths in the lattice lead to the single, globally maximal node WB, interpretation (1) of a non-rigid structure is accepted. We look no further for other nodes because there is no evidence in the display which suggests that interpretation (1) is indeed incorrect. This third alternative for adjudicating among may maximal nodes, where a committment is made to an interpretation of an earlier module, thus can create a Garden Path to a particular class of maximal nodes.

The situation changes, however, upon prompting. For example, if the rotation axis is altered to lie significantly off the frontal plane (then there is evidence against B), or if one is instructed to seek a completely rigid interpretation – which means faulting W and B right at the start. In this last case the lattice will then have the structure illustrated in Figure 7 (right) and a rigid interpretation will be recognized as maximal. We should not view these changes in the lattice structure as a failure in the assimilation process. Rather, the competence is there, for the complete lattice indeed will contain the "correct" interpretation[5], but performance limitations have led to "shortcuts", such as part-based decompositions, that are typically robust in the natural world.

5.0 Discussion

At the outset, we mentioned three different schemes for integrating modules. How does our lattice framework mesh with these alternatives?

5.1 Coupled Modules

Several studies have shown that ambiguities in interpretations reached by two separate modules can be removed by each constraining the other. Current examples include stereo and motion, shading and motion, and stereo and shading [3]. There are two ways in which these constraints may act.

In the simplest case, when different kinds of measurements are interpreted by the different modules, as in stereo and motion, then the additional data remove ambiguities simply by choosing the one interpretation common to both modules. In our lattice framework this would be equivalent to two (or more) alternative interpretations lying in a maximal node, and then new data (from another module) eliminates one (or more) of the alternatives.

A second case of coupling occurs when distinct, optional interpretations are available from a single module, as when there are two or more maximal nodes, each having non-overlapping premise labels. Now data from a second module, not previously introduced, may eliminate some of these maximal nodes. Such an elimination process would first involve constructing a new lattice, and then checking for inconsistencies that may be subject to the simultaneous application of constraints previously ignored. Figure 2 is a case in point: the addition of the fixed axis premise F changed the structure of the original RD lattice and led to a new maximal node containing a new interpretation.

5.2 Veto Schemes

Choosing a single interpretation by majority vote among the concerned modules makes no sense if inconsistent conclusions are not ruled out. But this then entails reasoning about the implications of the data in terms of world events, not in terms of whether a Stereo, SFM or Form module is "turned-on" or not. Once one admits to reasoning about what the data mean in terms of premises or constraints about the world, then in effect the viable nodes of our lattice are being used. Once these consistent interpretations are found, we simply propose placing a partial order on them, and picking as a preferred interpretation – or perception – one which corresponds to a local maximum within this partial order.

5.3 Probability Schemes

Our partial ordering of interpretations gives the premises a weight 0 or 1 – nothing in between, as conditional probability methods would like [19], [20], [46]. However, this does not mean that we can not generate a probability scheme using the same classes of premises and in which preferred means "locally more probable", thereby mapping any lattice into a probabilistic framework. Even if weights other than 0 or 1 are used, the most basic notion of our proposal remains: the preferred interpretation is associated with a maximal node in the lattice, which means there is no path to a higher node that carries a consistent interpretation of the data. Likewise, any probability scheme must rule out combinations of conditional statements or premises that are inconsistent. The assignment of weights won't by itself change the basic topology of the lattice of partial orderings. On the other hand, any probability scheme has a natural total ordering based on the modeled probability of any state. Hence probability schemes may sanction a stronger ordering than the simple one proposed here. For example, as discussed in relation to Figure 3, a probability analysis may conclude that the event $RF\overline{D}$ is much more likely than $\overline{RF}D$. In order to use this conclusion an extra arc needs to be added to the lattice, as in Figure 3.

5.4 Partial Orderings

The use of probabilities is not the only way that lattice orderings may be modified. In particular, many partial orderings can be placed on (classes of) interpretations by considering the preferences for accepting various premises. There is a minimal ordering in which the relative preferences of different premises is not considered; rather the ordering is entirely based on the preference of the positive form of a premise over its negation, given that all else is equal. This minimal ordering was used in our examples, and is appropriate in situations where the relative probabilites of various events is not known quantitively. In such an ordering we are simply assuming that the premise is more probable than its converse (in any situation it applies). Inferences based on this ordering can therefore be expected to be relatively robust under changes in the probability distributions of various events. Other orderings, for example ones which sanction preference relations between different premises (as appropriate for a veto scheme), can also be considered. However, in using such an ordering it is important to note what additional assumptions they are implicitly making about the scene and about the relative probabilities of different events. Here, we feel, an important bridge may be built to current knowledge representation formalisms, such as the ϵ-semantics of Pearl & Geffner [47], or the work on default reasoning using statistical knowledge by Bacchus [6].

6.0 "Take-Home" Messages

Although our assimilation proposal is primarily a competence theory, not an algorithm, the idea that conflicts between modules can be resolved by faulting premises about world structures does have implications for both biological and machine vision.

First, our proposal hinges on the notion that the task of a perceptual system is to explain its sense data as arising from world events, consistent with its models of the world. Our lattice theory provides a new framework for understanding some of the elements and rules that underly such a perceptual process. The key elements are the fallible premises about the world structure (not features and image relations!). One major rule is to choose interpretations that can explain our observations without unnecessarily faulting premises. Faulting these premises should be in the interest of obtaining consistent interpretations. Clearly such a search for interpretations that are consistent both with the premises and the sense data can not occur simply by having a list of all acceptable interpretations. Rather, some kind of reasoning process must occur, as depicted in part by our lattice. Thus, our proposal points to three tough problems for both biological and machine vision: (1) What constitute general yet powerful premises? (2) What constitutes consistency? (3) What are the rules of the process which reasons about consistency among the chosen premises, given the sense data?

For those engaged in both biological and machine vision, our proposal stresses the need to enter knowledge about the world early in information processing. Secondly, it will not be sufficient simply to use this knowledge as constraints that regularize the data and passively seek "optimal" solutions [10], [22], [31], [52], [58]. Instead, as emphasized especially by Gregory [24], [25], Helmoltz [27], Mackworth [38], [39], and Rock [55], the key to interpreting our sense data is to be able to reason about the consistency of an interpretation within the chosen conceptual models of the world (i.e. the premises). Discovering general, powerful premises will be a rewarding challenge. Somewhat of a surprise to us was that in order to understand the Ames Trapezoid illusion, we were forced to postulate a grouping of the display into "pre-objects" such as the window itself and the bar. Perhaps, in retrospect, this should be expected because, after all, we are proposing that the grouping of the image elements should be based upon premises about what's in the world, not about what's in the image [21], [36], [43], [50], [69]. Thus, the most important take-home message, easily lost sight of, is that seeing can be understood only when we grasp how knowledge about the world can be used to organize even the earliest elements of the visual information process [25], [44], [45], [63], [66], [68].

Acknowlegment

Useful comments on this or earlier material were given by Jacob Feldman and J. McClelland. The manuscript was prepared by William Gilson.

7.0 References

[1] J.K. Aggerwal and N. Nandhakumar, "On the computation of motion from sequences of images – a review," *Proc. IEEE*, vol. 76, pp. 9179̃35, 1988.

[2] J. Aloimonos, "Visual shape computation," *Proc. IEEE*, vol. 76, pp. 8999̃16, 1988.

[3] J. Aloimonos and D. Schulman, Integration of Visual Modules: An Extension of the Marr Paradigm. London: Academic Press, 1989. Also *Proc. 1989 DARPA IU Workshop*, p. 507-551.

[4] A. Ames, "Visual perception and the rotating trapexoidal window." *Psych. Monographs*, vol. 65, Whole no. 329, 1951.

[5] A. Ansaldi, L. DeFloriani, and B. Falcidiano, "Geometric modelling of solid objects using face-adjacency graph representation," *SIGRAPH*, vol. 19, p. 3, 1985.

[6] F. Bacchus, "Default reasoning using statistical knowledge," TR CS-90-39, Dept. of Comp. Sci., Univ. of Waterloo, Los Angeles, CA, 1988.

[7] R. Bajcsy and F. Solina, "Three dimensional object representation revisted," *Conf. Comp. Vision (London) IEEE-CS*, pp. 2312̃40, June 1987.

[8] D.H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, pp. 1111̃22, 1981.

[9] D.H. Ballard and C. Brown, *Computer Vision*. Englewood Cliffs, NJ: Prentice Hall, 1982.

[10] D.H. Ballard, G.E. Hinton, and T.J. Sejnowski, "Parallel visual computation," *Nature*, vol. 306, pp. 212̃6, 1983.

[11] A. Barr, "Global and local deformations of solid primitives," *Comp. Graphics*, vol. 18, pp. 213̃0, 1984.

[12] H.G. Barrow and J.M. Tennenbaum, "Recovering intrinsic scene characteristics from images," in *Computer Vision Systems*, A. Hanson & E. Riseman, Eds. New York: Academic Press, 1978.

[13] B. Bennett, D. Hoffman, and C. Prakesh, *Observer Mechanics.* London: Academic Press, 1989.

[14] I. Biederman, "Recognition by components." *Psych. Rev.*, vol. 94, pp. 1151̃47, 1987.

[15] T.O. Binford, "Inferring surfaces from images," *Art. Intell.*, vol. 17, pp. 2052̃44, 1981.

[16] A. Bobick, "A hybrid approach to structure from motion," in *Motion: Representation and Perception*, N. Badler and J. Tsotsos, Eds. New York: North-Holland, pp. 911̃09, 1986.

[17] M. Brady and A. Yuille, "Inferring 3D orientation from 2D contour: an extremum principle," in W. Richards, *Natural Computation.* Cambridge, MA: MIT Press, pp. 99˜106, 1988.

[18] H.H. Bülthoff and H.A. Mallot, "Integration of depth modules: stereo and shading," *Jrl. Opt. Soc. Am. A*, vol. 5, pp. 1749˜1758, 1988.

[19] P. Cheeseman, "An inquiry into computer understanding," *Comput. Intell.*, vol. 4, pp. 58˜66, 1988.

[20] J.J. Clark and A.L. Yuille, *Data Fusion for Sensory Information Processing Systems.* Boston: Kluwes Academic, 1990.

[21] J. Feldman, "Perceptual decomposition as inference: continuous curvilinear processes," M.S. thesis, Dept. Brain and Cognitive Sciences, MIT, Cambridge, MA, 1990.

[22] J.A. Feldman and D.H. Ballard, "Connectionist models and their properties," *Cog. Sci.*, vol. 6, pp. 205˜254, 1982.

[23] J. Gibson, *The Ecological Approaches to Visual Perception.* Boston: Houghton-Mifflin, 1979.

[24] R.L. Gregory, *The Intelligent Eye.* New Jersey: McGraw Hill (eg. p. 31), 1970.

[25] R.L. Gregory, "Perceptions as hypotheses," in *The Psychology of Vision*, H.C. Longuet-Higgins and N.S. Sutherland, Eds. London: The Royal Society, pp. 1371̃49, 1980.

[26] W.E.L. Grimson, "Binocular shading and visual surface reconstruction," *Comp. Vis. Graphics Image Proc.*, vol. 28, pp. 194̃3, 1984.

[27] H. Helmholtz, *Handbook of Physiological Optics.* Dover reprint of 1925 edition ed. by J.P.C. Southall, 3 volumes, 1963.

[28] D. Hoffman and B. Bennett, "The computation of structure from fixed-axis motion: rigid structures," *Biol. Cybern.*, vol. 54, pp. 718̃3, 1986.

[29] D.D. Hoffman and B.E. Flinchbaugh, "The interpretation of biological motion," *Biol. Cybern.*, vol. 42, pp. 1952̃04, 1982.

[30] D. Hoffman and W. Richards, "Parts of recognition," *Cognition*, vol. 18, pp. 659̃6, 1984.

[31] J.J. Hopfield and D.W. Tank, "Neural computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, pp. 1411̃52, 1985.

[32] J. Hochberg, "Machines should not see as people do, but must know how people see," *Computer Vision, Graphics and Image Proc.*, vol. 37, pp. 2212̃37, 1987.

[33] R. Jain, Assimiliation Workshop, Univ. Mich. Ann Arbor, June 1990.

[34] R. Jain, "Environment models and information assimilation," Technical Report, IBM Almaden Research Labs, June 1989.

[35] A. Jepson and W. Richards, "Perception and perceivers," Technical Report, Dept. of Computer Science, University of Toronto, 1990. (Also presented in part at the University of Minnesota Workshop on Vision and 3D Representation, May 1989.)

[36] T. Kanade, "Recovery of the three-dimensional shape of an object from a single view," *Art. Intell.*, vol. 17, pp. 409-460, 1981.

[37] D. Lowe, *Perceptual Organization and Visual Recognition.* Boston, MA: Kluwer Academic, 1985.

[38] A.K. Mackworth, "Recovering the meaning of diagrams and sketches," *Proc. Graphics Interface '83*, Edmonton, pp. 3133̃17, 1983.

[39] A.K. Mackworth, J. Mulder, and W.S. Havens, "Hierarchical arc consistency: exploiting structured domains in constraint satisfaction problems," *Computational Intell.*, vol. 1, no. 3, pp. 1181̃26, 1985.

[40] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: W.H. Freeman, 1982.

[41] D. Marr, D. and K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proc. Roy. Soc. Lond. B*, vol. 200, pp. 2692̃94, 1978.

[42] J.L. Marroquin, S.K. Mitter, and T. Poggio, "Probabilistic solutions of ill-posed problems in computational vision," *J. Am. Stat. Assoc.*, vol. 82, pp. 768̃9, 1987.

[43] E. Mjolsness, G. Gindi, and P. Anandan, "Optimization in model matching and perceptual organization," *Neural Computation*, vol. 1, pp. 2182̃29, 1989.

[44] D. Mumford, "On the computational architecture of the neocortex I. The tole of the thalamus-cortical loop," *Biol. Cybern.*, vol. 65, pp. 135-145.

[45] K. Nakayama and S. Shimojo, "DaVinci stereopsis: depth and subjective occluding contours from unpaired image points," *Vis. Res.*, vol. 30, pp. 1811-1825, 1990.

[46] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Matteo: Morgan Kaufman, 1988.

[47] J. Pearl and H. Geffner, "Probabilistic semantics for a subset of default reasoning," TR R-93-III, Cognitive Systems Lab., Univ. of Calif., Los Angeles, 1988.

[48] A. Pentland, "Shading with texture," *Art. Intell.*, vol. 29, pp. 1471̃70, 1986.

[49] A. Pentland, "Part segmentation for object recognition," *Neural Computation*, vol. 1, pp. 829̃1, 1989.

[50] A. Pentland, "Perceptual organization and the representation of natural form," *Art. Intell.*, vol. 28, pp. 2933̃31, 1986.

[51] T. Poggio, E.B. Gamble and J.J. Little, "Parallel integration of visual modules," *Science*, vol. 242, pp. 436-440, 1988.

[52] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 3143̃19, 1985.

[53] R. Reiter and A. Mackworth, "A logical framework for depiction and image interpretation," *Art. Intell.*, vol. 41, pp. 125-155, 1989. (Also "The logic of depiction," UBC Dept. Computer Science Tech. Report 87-24, 1987.)

[54] W. Richards, "Structure from stereo and motion," *Jrl. Opt. Soc. Am. A*, vol. 2, pp. 3433̃49, 1984.

[55] I. Rock, *The Logic of Perception.* Cambridge, MA: Bradford (MIT Press), 1983.

[56] A. Rosenfeld, "Computer vision: basic principles," *Proc. IEEE*, vol. 76, pp. 8638̃68, 1988.

[57] J. Rubin and W. Richards, "Visual perception of moving parts," *Jrl. Opt. Soc. Am. A*, vol. 5, pp. 20452̃049, 1988.

[58] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations.* Cambridge, MA: MIT Press, 1986.

[59] K.A. Stevens, "The visual interpretations of surface contours," *Art. Intel.*, vol. 17, pp. 174̃5, 1981.

[60] K.A. Stevens, "Line of curvature constraint and the interpretation of 3D shape from parallel surface contours," in *Natural Computation*, W. Richards, Ed. Cambridge, MA: MIT Press, pp. 1071̃14, 1988.

[61] D. Terzopoulos, "Integrating visual information from multiple sources," in *From Pixels to Predicates*, A. Pentland, Ed. Norwood, NJ: Ablex, pp. 1111̃42, 1986.

[62] S. Truvé, “Toward a specification language for physical objects,” Technical Report, Programming Methodology Group, Dept. Computer Science, Chalmers, Göteborg, Sweden. (See also Chapter 11 in *Natural Computation*, W. Richards, Ed. Cambridge, MA: MIT Press, 1987.

[63] J.K. Tsotsos, “A complexity level analysis of vision,” *Behavioral and Brain Sciences*, vol. 13-3, pp. 4234̃55, 1990.

[64] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, MA: MIT Press, 1979.

[65] S. Ullman, “Maximizing rigidity: the incremental recovery of 3D structure from rigid and non-rigid motion,” *Perception*, vol. 13, pp. 2552̃74, 1980.

[66] S. Ullman, “Sequence seeking and counter streams: a model for information processing in the cortex,” MIT A.I. Memo 1311, 1992.

[67] A. Waxman and J.H. Duncan, “Binocular image flows: steps toward stereo-motion fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-8, pp. 7157̃29, 1986.

[68] A. Witkin and J.M. Tennenbaum, “What is perceptual organization for?” *Proc. IJCAI-83*, pp. 10231̃026, 1983.

[69] A. Witkin and J.M. Tennenbaum, "On the role of structure in vision," in *Human and Machine Vision*, J. Beck, B. Hope, and A. Rosenfeld, Eds. New York: Academic, 1983.

Footnotes for manuscript SMC 091-08-0823, "A Lattice Framework for Integrating Vision Modules", by Allan Jepson & Whitman Richards.

Allan Jepson was supported by NSERC Canada. Whitman Richards was supported by AFOSR 89-504. Allan Jepson is with the Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Ontario M5S 1A7. Whitman Richards is with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 79 Amherst St., Cambridge, Massachusetts 02139.

[1] For clarity, we treat here the data received by a module as the sense data. However, in a hierarchical arrangement of modules, the interpretation offered by one module can be the "input data" for the next module in the information processing pathway. See [13] for the original proposal and formal treatment of this idea.

[2] Clearly here the intention is to state the premises such that the positive version is preferred over its negation.

[3] In order for a unique maximum to be generated, then the two children of $\overline{R}D$ (Figure 1), namely $\overline{RX}D$ and $\overline{R}X\overline{D}$, must both be inconsistent. But this is impossible as long as $\overline{R}D$ is itself consistent. Similarly for $R\overline{D}$.

[4] When configurations such as these rotate with axes significantly out of the frontal-parallel plane, however, there is evidence for additional structure and hence the lattice of Figure 6 will change.

[5] Although the correct interpretation may not appear in a maximal node there will always be at least one maximal node that contains a rigid interpretation.

Figure captions for manuscript SMC 091-08-0823, "A Lattice Framework for Integrating Vision Modules", by Allan Jepson & Whitman Richards.

Fig. 1. A simple fault-lattice based upon a rigid motion premise R and a stereo disparity premise D and their faults. Cross-hatching indicates an inconsistent node in the lattice. $\overline{R}$ and $\overline{D}$ indicate faulted premises. The interpretation associated with a node is indicated beneath that node.

Fig. 2. An augmentation of the fault lattice of Figure 1 by a new premise "X". Premises R and D are repectively "rigid motion" and "planar stereo disparity" as before. Premise "X" is assumed not to invalidate the RD inconsistency, so two nodes are immediately deleted as inconsistent, leaving both S and M maximal nodes.

Fig. 3. The fault lattice for R, F and D, where premises R and D are "rigid motion" and "planar disparity" as before, and F is a fixed-axis premise. The display is assumed to be a flat TV screen. There are still two maximal nodes as indicated by the arrows, but now the non-rigid, planar interpretation (stereo) has more faulted premises.

Fig. 4. Three possible schemes for altering the lattice of Figure 3 in order to obtain a unique maximal node. Top: (A) The voting scheme simply counts faults and allows a path to higher nodes with fewer faults (dashed arrow). Middle: (B) Adding weights to the premises can move a node with only one valid premise to a maximal position. Bottom: (C) The addition of another premise, here associated with the concept "picture" can create a globally consistent maximal node. This solution assimilates by explaining the data in a manner consistent with all the active premises.

Fig. 5. The Ames Trapezoid Window configuration plus bar is illustrated from one vantage point in the middle panel (the circles simply highlight feature points). The "structure-from-motion" module captures the major feature points (indicated by the circles) and tests for a rigid 3D configuration. Similarly, the "form" module extracts two parts – the window and the bar – and computes a 3D orientation for these separate parts.

Fig. 6. The fault lattice for the three premises R, W and B used in the evaluation of the image data presented as the Ames Trapezoid display. The numbers refer to plausible, consistent interpretations of the data that are associated with a node in the fault-lattice. The maximal node ($\overline{R}WB$) contains a non-rigid interpretation (1) as indicated by the faulted R. The rigid interpretation (2) lies in a second maximal node ($R\overline{WB}$) with two faults.

Fig. 7. An explanation for the Ames illusion based upon a parts-based strategy for image-interpretation. If the window and bar are first identified and then the consequences of premises W and B are explored, a unique maximal node (WB) will be found that contains the plausible (non-rigid) interpretation (1). The search is then terminated. On the other hand, if evidence for the window and bar being separate parts is removed, such as by presenting only the feature points (circles in Figure 5), then the entire configuration is treated as one grouping and a consistent rigid interpretation (2) is found in a ($R\overline{WB}$) node.

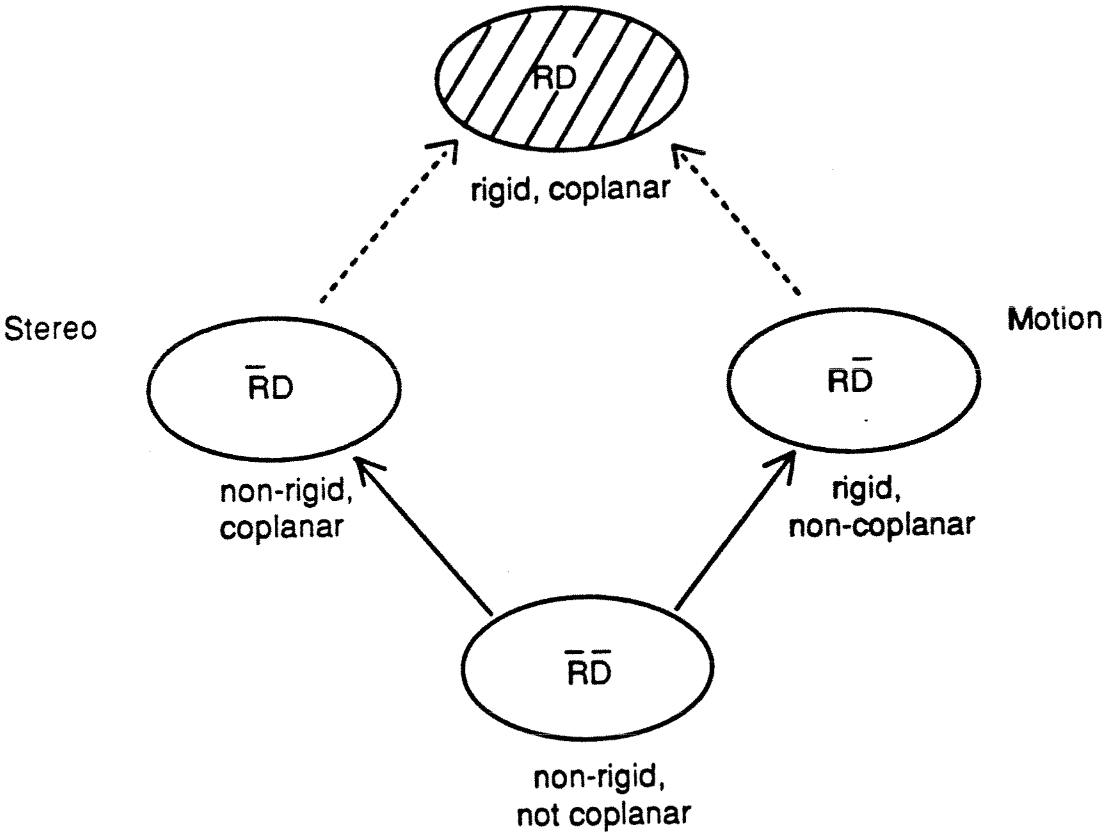

RD
rigid, coplanar
Stereo
$\bar{R}D$
non-rigid, coplanar
Motion
$R\bar{D}$
rigid, non-coplanar
$\bar{R}\bar{D}$
non-rigid, not coplanar

Fig 1 Assim

Figure 1

Jepson & Richards

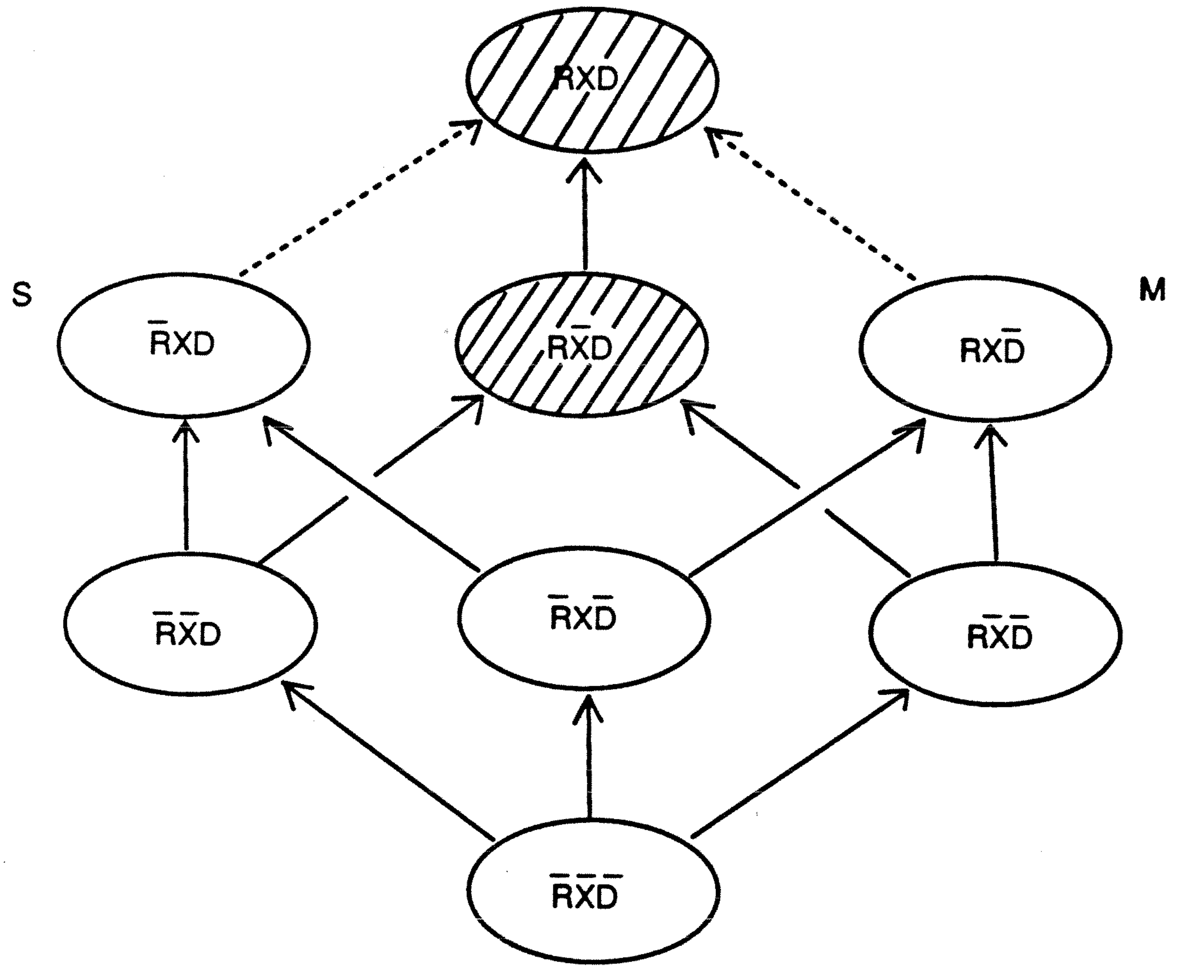

RXD
S
$\bar{R}XD$
$R\bar{X}D$
$RX\bar{D}$
M
$\bar{R}\bar{X}D$
$\bar{R}X\bar{D}$
$R\bar{X}\bar{D}$
$\bar{R}\bar{X}\bar{D}$

Fig 2 Assim

Figure 2

Jepson & Richards

SMC 091-08-0823

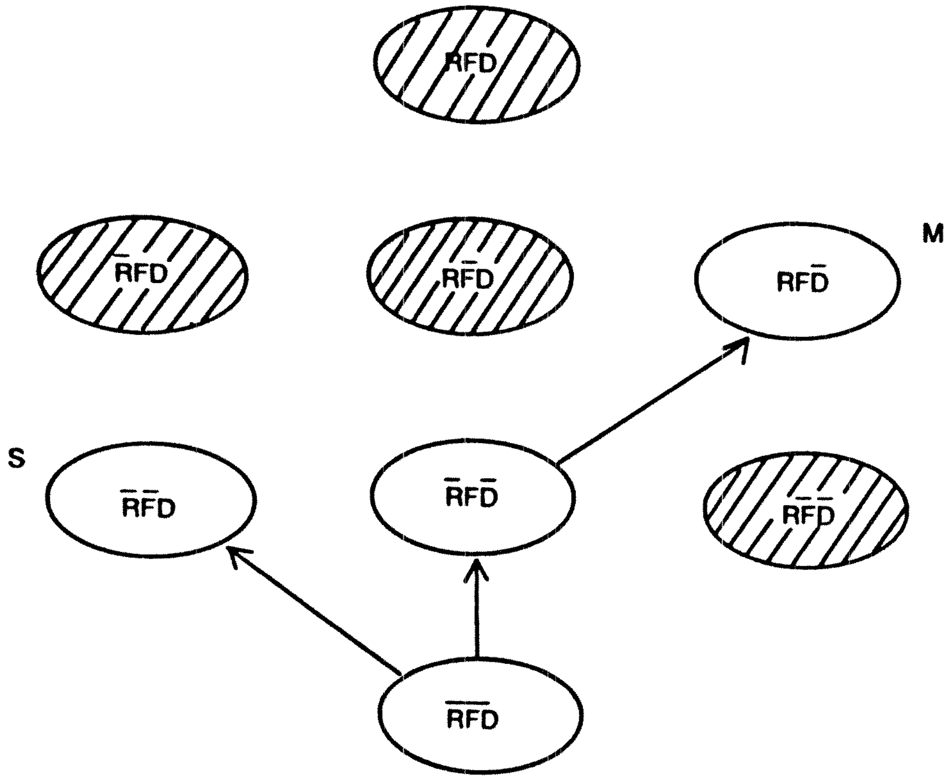


Fig 3A Assim

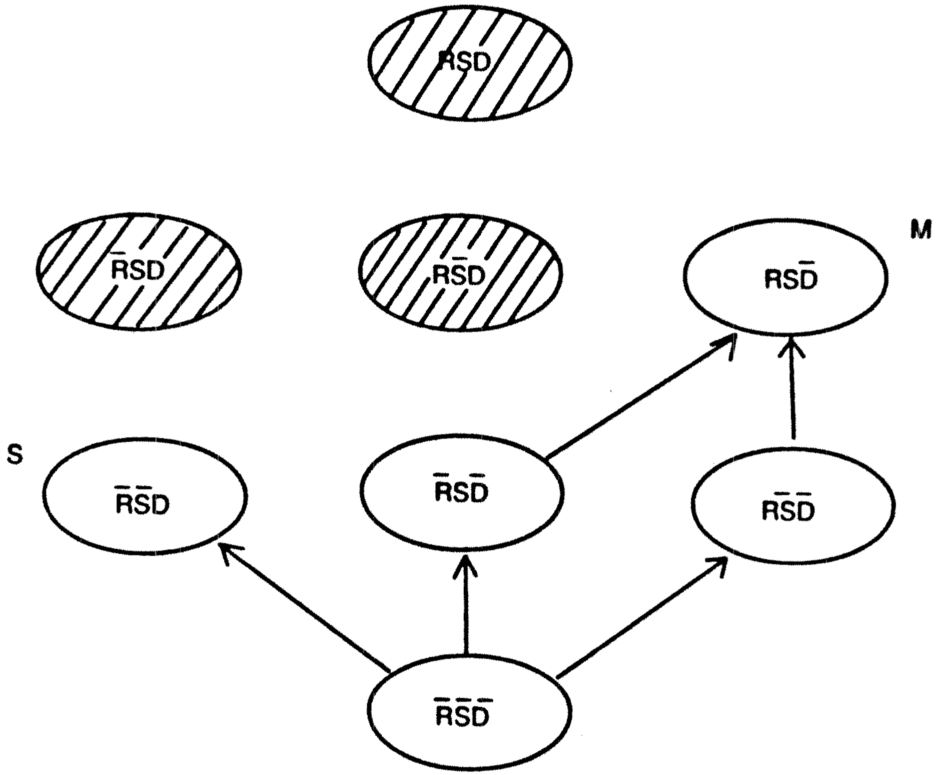
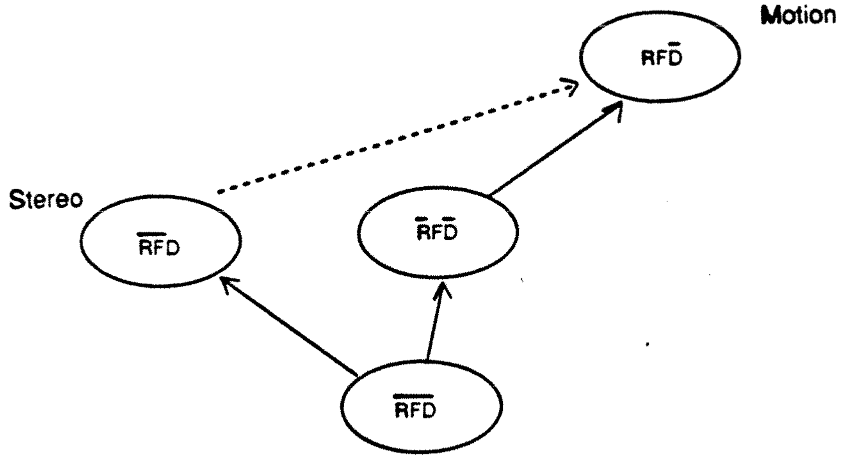


Fig 3B Assim

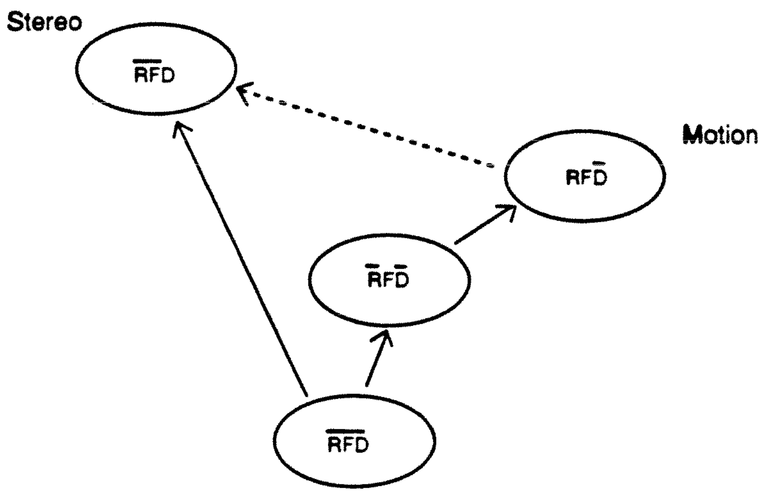
Figure 3

Jepson & Richards

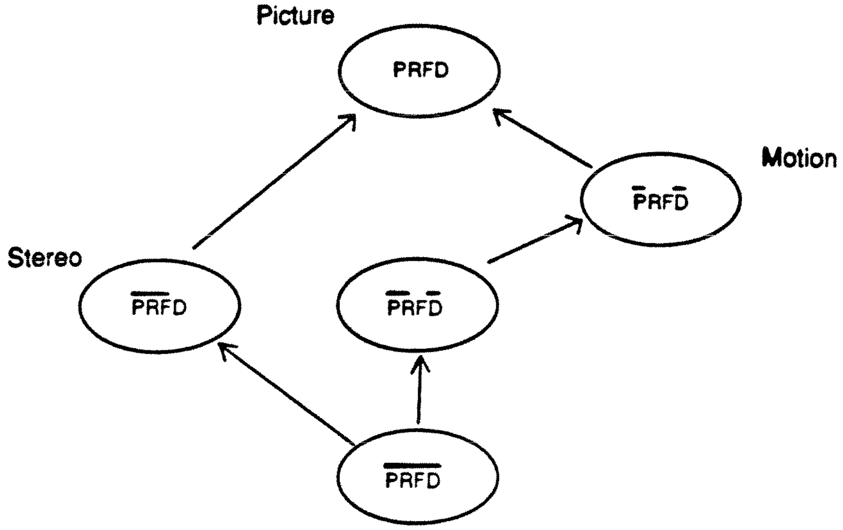
SMC 091-08-0823



VOTING



"WEIGHTS"



ASSIMILATION by CONCEPT

FIGURE 4

Figure 4

Jepson & Richards.

SMC 091-08-0823

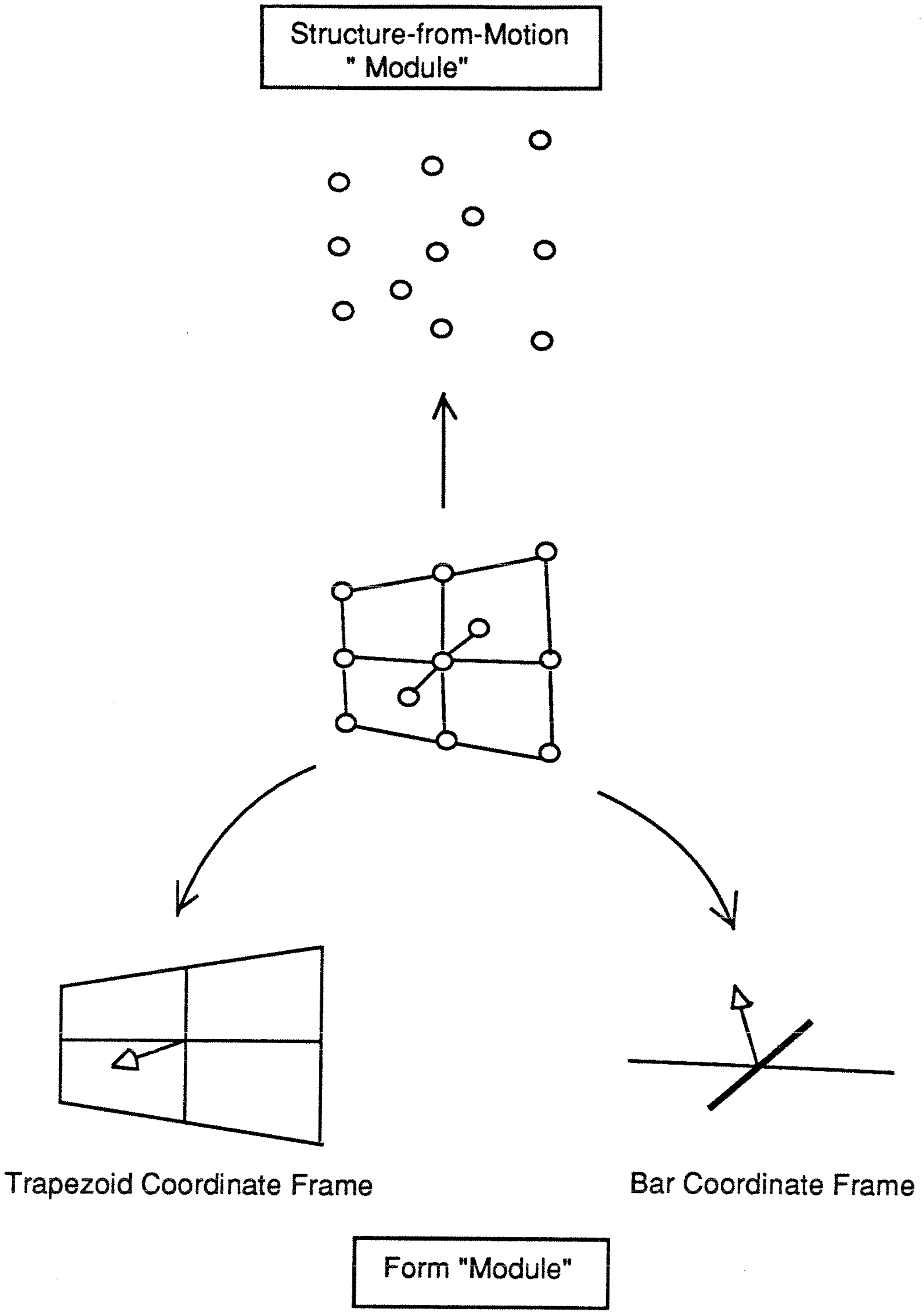

Structure-from-Motion
" Module"
Trapezoid Coordinate Frame
Bar Coordinate Frame
Form "Module"

Fig 5 Assim

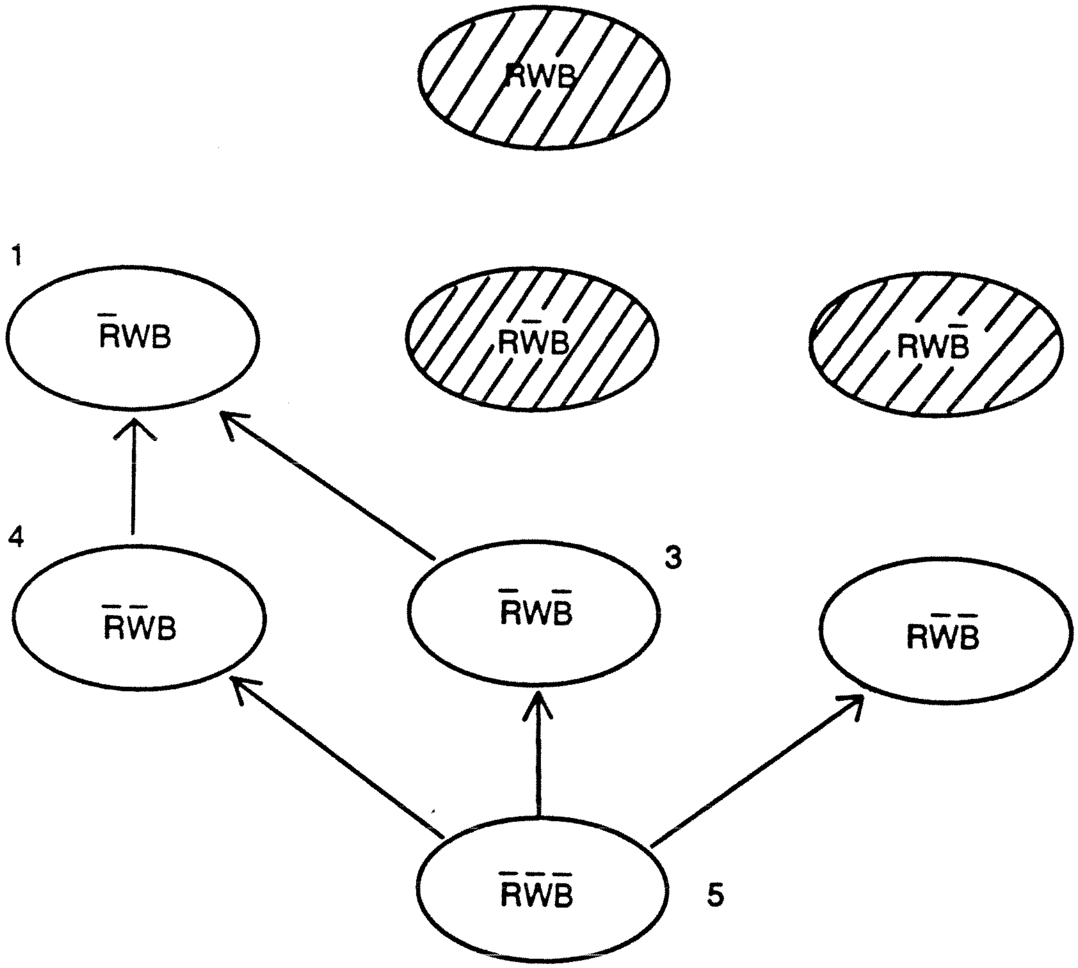

RWB
1
$\overline{R}WB$
$R\overline{W}B$
$RW\overline{B}$
4
$\overline{R}\overline{W}B$
3
$\overline{R}W\overline{B}$
2
$R\overline{W}\overline{B}$
$\overline{R}\overline{W}\overline{B}$
5

Fig 6 Assim

Figure 6

Jepson & Richards

SMC 091-08-0823

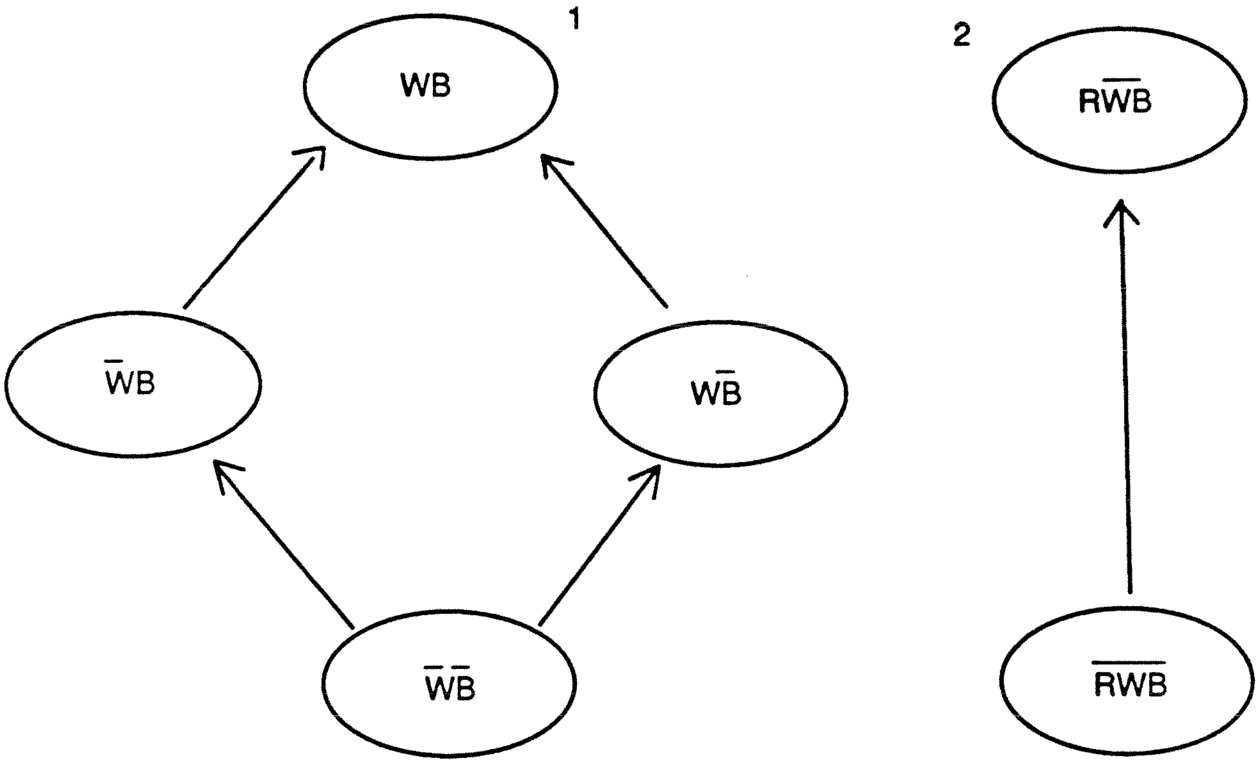

1
WB
$\overline{W}B$
$W\overline{B}$
$\overline{W}\overline{B}$
2
$R\overline{W}B$
$\overline{RWB}$

Fig 7 Assim

Figure 7

Jepson & Richards

SMC 091-08-0823