CVGIP: IMAGE UNDERSTANDING
Vol. 54, No. 2, September, pp. 184–205, 1991

A Qualitative Approach to Dynamic Scene Understanding

BIR BHANU

College of Engineering, University of California, Riverside, California 92521

AND

WILHELM BURGER

Johannes Kepler University, Linz, Austria

Received June 25, 1987; accepted March 9, 1990

Vision systems for mobile robots are required to handle complex dynamic scenes. Vehicle motion and individually moving objects in the field of view contribute to a continually changing camera image. The goal of "dynamic scene understanding" is to find consistent explanations for all changes in the image in terms of three-dimensional camera motion, individual object-motion, and static scene structure. We describe a new approach to this problem which departs from previous work by emphasizing a qualitative line of reasoning and modeling. We have extended the original Focus-of-Expansion concept to the so-called Fuzzy FOE, where we do not compute a singular point in the image, but a connected image region that marks the approximate direction of heading. A rule-based reasoning engine analyzes the resulting "derotated" displacement field for certain events and incrementally builds a three-dimensional Qualitative Scene Model. This model comprises a collection of scene hypotheses, each representing a feasible and distinct interpretation of the current scene. This paper focuses on this qualitative approach for dynamic scene understanding. Examples are given for synthetic as well as for real outdoor image sequences. © 1991 Academic Press, Inc.

1. INTRODUCTION

Vision is an indispensable source of information for the operation of mobile robots or autonomous vehicles. Even when robots are equipped with accurate inertial navigation systems, the accumulation of position errors requires periodic corrections. The execution of mission tasks involving search, exploration, or manipulation in particular appear almost impossible without visual support. While the robot is moving, the resulting images acquired by its camera are changing continually, even if the observed environment is completely static. In this case, image motion can be used to obtain useful information about the robot's self-motion and about the three-dimensional layout of the scene, commonly referred to as "motion stereo."

In dynamic environments, the potential appearance of individually moving objects in the scene adds another level of complexity. Object motion and camera motion interfere and moving objects may not even cause any image motion at all. The environment cannot be treated as a single rigid object but possibly as several of them, one of which must serve as a global reference. Any change observed in the 2-D image is always the result of a change in 3-D space, caused either by self-motion or by individual object motion. Finding consistent interpretations for every change in the image in terms of self-motion, 3-D scene structure, and object motion is the objective of "dynamic scene understanding."

Previous work in motion understanding has concentrated on numerical approaches for the reconstruction of 3-D motion and scene structure from 2-D image sequences. In the traditional approach, structure and motion of a rigid object are computed simultaneously from successive perspective views by solving systems of linear or nonlinear equations [1–5]. This technique is reportedly noise sensitive even when more than two frames are used [6, 7]. Nonrigid motion, or the presence of several moving objects in the field of view, may produce a relatively large error for the final solution to the system of equations. However, due to the inherent ambiguities in motion analysis, an acceptable "rigid" solution may be found even when parts of the scene are actually moving in 3-D. Thus there are cases where the movements of individual entities in the field of view cannot be detected by the classic scheme. Adiv [8] generalized this approach to handle scenes with multiple moving objects, using an iterative grouping process to segment the optical flow field.

For applications with mainly translational camera movements, such as land vehicles, alternative approaches have been developed to make use of this particular form of self-motion [9–11]. An important concept

1049-9660/91 $3.00

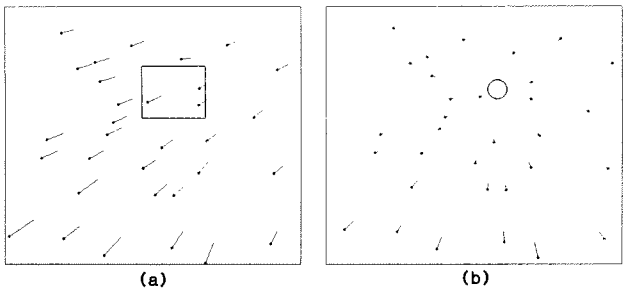

FIG. 1. A typical displacement field obtained from a moving camera undergoing translation and rotation. (a) The search area for the *Focus of Expansion* is marked by a square in the original (simulated) displacement field. (b) The *derotated* displacement field with the approximate location of the FOE marked by a circle.

related to this class of techniques is the *Focus of Expansion* (FOE), i.e., the image location from which all points seem to diverge radially under pure (forward) camera translation. Figure 1 shows a typical set of displacement vectors obtained from a moving camera which undergoes translation and rotation. The FOE points in the direction of vehicle translation between two consecutive images.

Given the FOE for a pair of images, the 3-D distance of any (static) environmental point can be found easily from its *velocity* of radial divergence. In practice, locating the FOE *accurately* is difficult or even impossible under arbitrary camera motion or noisy conditions. Consequently, planar motion or even pure camera translation has been assumed by other researchers in this field [12–15].

To employ the FOE technique for our task, we have extended the original concept to compute only an *approximation* of the FOE for almost arbitrary camera motion. This so-called *Fuzzy FOE* is not specified by a single point (pixel) in the image, but by a connected image *region* that marks the approximate direction of heading [16]. Besides the computational problems involved in computing the FOE precisely, there appears to be also a psychological motivation for the Fuzzy FOE. Under comparable conditions (i.e., observer rotation), even human subjects reportedly have difficulties in estimating the exact direction of heading [17].

While it has been common to take the scene structure as a *by-product* of the rigid motion computation, we argue that the existence of an internal 3-D model of the scene is a necessary *prerequisite* for motion detection and analysis. Given the location of the FOE and a purely translational displacement field, some forms of 3-D motion are easy to detect, whereas others require more sophisticated reasoning steps. For example, an image feature moving *toward* the FOE is a striking evidence for 3-D motion in the scene; in particular, something must be moving into the camera's current trajectory.

A more subtle case is given in Fig. 2, which shows two successive frames as the camera approaches an intersection. The shaded area in the center represents the approximate location of the FOE (i.e., the Fuzzy FOE). Two points of interest are tracked, one located on the truck (A) and the other on the building (B). The point on the building (B) diverges away from the FOE at some rate and can thus be interpreted as static in 3-D at some finite distance. However, the point on the truck (A) stays at a *constant* image location and could therefore also be static, though at *infinite* distance. The actual motion of the truck, which could potentially collide with the camera, would remain undetected. Although in reality the truck will probably cause some image motion, noise will still make it difficult to assess. Additional information about the spatial layout of the scene is necessary to resolve these ambiguities. Assuming that we had *occlusion* as another source of information, the reasoning process could be like this:

> *"Point B is diverging from the FOE and thus lies at finite distance and is static. Point A, if static, would be an infinite distance. Since the object of point A is occluding the object associated with point B, A must be closer than B. From this contradiction we conclude that A must be moving. . . ."*

Instead of occlusion analysis, we have actually employed simple heuristics about the scene structure for additional 3-D clues.

The main elements of our approach are illustrated in Fig. 3 which shows different levels of data and the processes which operate on them. Images are treated pair after pair and we assume that displacement vectors between corresponding points in successive images have been obtained by the correspondence process. Image features are given unique labels and tracked from frame to frame. Using the displacement vectors we compute the Fuzzy FOE and remove the effects of possible camera rotations. Consequently, the resulting "derotated" displacement field reflects the pure translation component of the camera motion. Note that up to this level, the flow of data is purely bottom-up. There is, however, control information supplied by the higher levels, such as the set of reference points which is believed to be stationary and

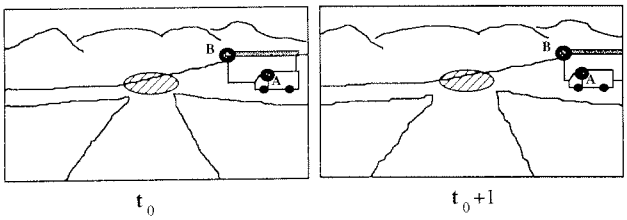

FIG. 2. Moving truck example. The camera is approaching an intersection. Two point features are tracked in the image: point A on the (moving) truck, and point B on the static building. The shaded area in the center represents the Fuzzy FOE. The static part of the image seems to expand from the FOE, while the truck, being on a collision path, stays at a constant image location.

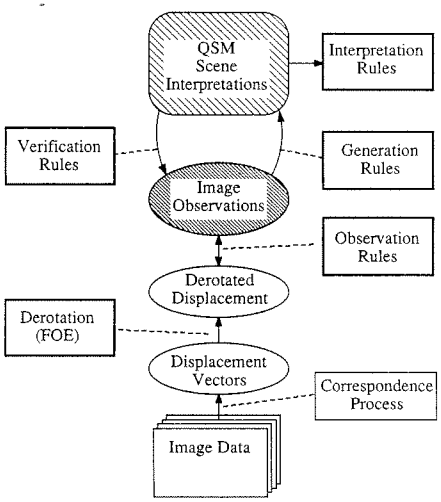

FIG. 3. Overall structure of the interpretation process and flow of data for the construction of the *Qualitative Scene Model (QSM)*. From the original displacement vectors (obtained by matching corresponding features), the *Fuzzy FOE* and the "derotated" displacement field are computed. *Observation rules* analyze the derotated displacement field for configurations and significant changes in the image in the context of motion understanding. QSM is built in a hypothesize-and-test cycle. *Generation rules* search for significant image events and place immediate conclusions (hypotheses) in the model. *Verification rules* check existing hypotheses for consistency with the changes occurring in the image. *Interpretation rules* assemble complete interpretations from partial interpretations. A set of environmental entities that are believed to be stationary is supplied by the QSM for use in the FOE-computation.

may be used to compute the FOE. Next, we use the 2-D locations and motion of points relative to each other and with respect to the Fuzzy FOE in the derotated displacement field to reason about the 3-D scene structure and independent object motion. Given only an approximate location of the FOE, *qualitative* properties of the displacement field are the main source of reasoning.

This process incrementally builds a model of the environment, in which the scene is again described in qualitative terms, such as the relative distances of features or how they move in 3-D space. The model comprises a collection of scene hypotheses, each representing a feasible and distinct interpretation of that scene. In particular, the model labels those features that are believed to be part of the static environment and are used as references for computing the FOE.

The *observation process* analyzes the derotated displacement field for configurations and changes in the image which are deemed significant in the context of motion understanding. A group of forward-chained rules extract and formulate those changes in the image in the form of *trigger events,* which are then checked for their consequences upon the current state of the model. On the other hand, information about the image is delivered "on demand," i.e., when it is needed to complete a reasoning step at a higher level of the reasoning process. This mechanism is implemented with goal-driven, backward-chaining rules.

The *Scene Interpretations*, which are the actual core of the *Qualitative Scene Model* (QSM), are hypotheses about the relationships between the *facts* found in the image and their meaning in 3-D space. The reasoning process which forms the scene interpretations has access to 2-D information in the form of (already abstracted) image observations. Two forms of processes (rules) contribute to the core model in a *generate-and-test* strategy. *Generation Rules* take newly created image observations (by forward chaining) and determine their consequences with respect to the current state of the model. *Verification Rules* attempt to check existing hypotheses in the model for their validity with respect to specific changes in the image. Naturally, verification relies heavily on a backward-chained part of the observation rules. QSM may contain multiple scene interpretations at the same time. Individual interpretations, however, are not kept as separate constructs inside this model, but they generally share their components (partial interpretations) among each other. It is the task of the *Interpretation Process* to assemble complete interpretations from partial interpretations, to rank them, and make results available to other reasoning processes.

In the following we concentrate on the qualitative reasoning and modeling aspects of our approach. Details on computing the Fuzzy FOE can be found elsewhere [16]. Section 2 describes the structure of the Qualitative Scene Model and how it is updated. Examples of the rules used in the reasoning engine are given in Section 3. This is followed by experiments on synthetic and real images in Section 4. Finally, Section 5 presents the conclusions of the paper.

2. A QUALITATIVE SCENE MODEL (QSM)

2.1. *Entities and Relations*

The Qualitative Scene Model (QSM) is a 3-D camera-centered interpretation of the scene that is built incrementally from the visual information gathered over time. The model is declarative and describes the status and behavior of its elements and the relationships between them in coarse, qualitative terms. No attempt is made to derive a precise geometric description of the scene in terms of 3-D structure and object motion.

The basic elements of the QSM are called *entities,* which are the 3-D counterparts of the 2-D *features* observed in the image. For example, the point feature A located in the image at x, y at time t, denoted by (FEATURE A t x y), has its 3-D counterpart in the model as (MEMBER A). Properties of entities and relationships between entities are expressed by assertions. For example, (STATIONARY 1) means that entity 1 is considered stationary (i.e., not moving) in the corresponding scene

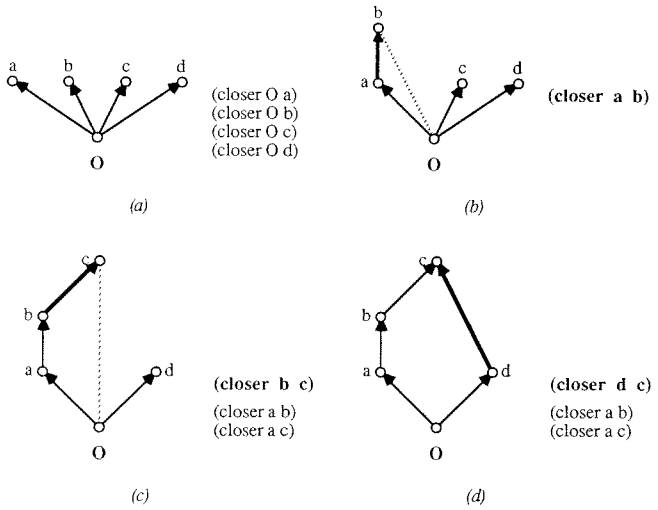

FIG. 4. Partial ordering in depth. (a) Initially, the relative depth is unknown for the entities a, b, c, d. They are only known to be in front of the image plane O. (b) (CLOSER a b) has been determined and is added to the list of facts. (c) (CLOSER b c) has been determined, which implies (CLOSER a c) by transitivity. (d) (CLOSER d c) has been determined; note that at this point nothing can be said about the relative depth between (a, d) and (b, d).

interpretation. In any scene interpretation, the current set of entities is divided into *stationary* (i.e., static) entities and *mobile* (i.e., possbily moving) entities.

2.2. *Modeling Static Scene Structure*

The static scene structure is modeled in the QSM in a fashion very similar to a camera-centered depth map. At time t, the 3-D location of any entity k with respect to the camera is completely specified by its image coordinates $x(k, t)$, $y(k, t)$ and its distance from the focal plane $z(k, t)$. However, in contrast to a regular depth map, the distance $z(k, t)$ is not represented by some numeric value, but by a qualitative spatial relationship between entities. In particular, the relation (CLOSER A B) means that entity A is believed to be *closer* to the camera than entity B in 3-D space. This relationship can be determined efficiently and reliably from the divergence of displacement vectors. While a regular depth map must be updated after every frame, this *semitopological* map requires no repetitive modifications as the camera moves forward through its environment. During this time, however, the model is continually refined as more *closer* relationships become evident (see Fig. 4).

2.3. *Modeling Object Motion*

Object motion is described at progressive levels of detail. The least that can be said about a moving entity C is (MOBILE C), which simply means that this entity is not part of the static environment. Once an entity has been identified as being in motion, it is considered *mobile* in all subsequent frames, even when its 3-D motion can no longer be verified.

Relative motion between two entities in 3-D may be detectable before the individual motion of a single entity becomes apparent. The fact (MOVEMENT-BETWEEN C D t) states that relative motion between C and D at time t has been concluded, but it tells nothing about which of the two entities are actually moving. This would be expressed by the more specific fact (MOVES C t) or (MOVES D t).

Details about how an entity moves within the camera-centered coordinate frame are expressed by additional facts, e.g., (MOVES-LEFT C t), (MOVES-DOWN C t), (*APPROACHING* C t), or (RECEDING C t).

2.4. *Interpretation Graph (IG)*

The QSM is structured as a directed graph whose nodes contain "partial scene interpretations." Each partial interpretation stands for a hypothesis represented by a collection of consistent assertions. Every node of this *interpretation graph* (IG), except the single root node, inherits the facts valid in its parent node(s). The root node itself holds all the facts that are globally true and is thus valid in any existing interpretation. Figure 5 shows a simple IG for a scene with 4 entities that are all believed to be stationary (the default assumption). Notice that at this point there exists no *complete interpretation*, i.e., a single node that contains (or inherits) a classification for every entity in the scene. Later, we shall show how complete interpretations are created by merging partial interpretations.

Fortunately, updating the QSM can be accomplished locally on partial interpretations and does not require complete interpretations. Assume, for example, that some rule has determined from the displacement field that two entities (*1*, *2*) must be moving relative to each other in 3-D, but could not determine which one was moving. This observation would lead to the new fact (MOVEMENT-BETWEEN *1* *2*) which is globally true and therefore asserted at the root node of the IG (Fig. 5). The model must now be updated to eliminate any interpretation that considers *both* entities *1* and *2* stationary, as accomplished by the following pair of rules:[1]

```
(defrule RELATIVE-MOTION-X
    (MOVEMENT-BETWEEN ?x ?y)
    (STATIONARY ?x)
  =>
```

[1] Here we use the actual syntax of *ART* [18] for defining rules:

```
(defrule RULE-NAME (premise-1) (premise-2) . . .
 => (action-1) (action 2) . . .)
```

Variables of the form ?A in premises and actions indicate local bindings within rules.

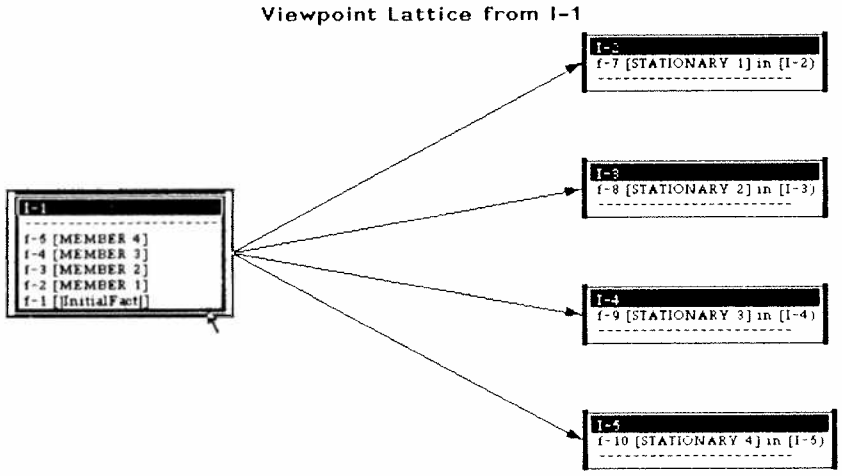

FIG. 5. Simple interpretation graph. The model contains four entities (1, 2, 3, 4), which are initially assumed to be stationary (by default). They are listed as members of the current model in the root node (left). For each entity, a partial interpretation has been created (represented by a child node) in which (STATIONARY *n*) is true.

```
    (assert (MOBILE ?y)))
(defrule RELATIVE-MOTION-Y
    (MOVEMENT-BETWEEN ?x ?y)
    (STATIONARY ?y)
  =>
    (assert (MOBILE ?x)))
```

A verbal interpretation of the first rule should help to clarify the syntax of these definitions: "*If a 3-D movement has been observed between entities X and Y then, assuming that X is stationary, entity Y must be mobile.*"

Trying to fire these rules, the system searches for the least specific partial interpretation (i.e., the node closest to the root) where all the premises are satisfied. By definition, the rule will then put the new assertions into this particular node. In Fig. 6 rule RELATIVE-MOTION-*X* fires in node I-2, asserting the new fact (MOBILE *2*) at this node. Note that the fact (MOVEMENT-BETWEEN *1 2*) is inherited from the root node. Similarly RELATIVE-MOTION-*Y* fires in node I-3, asserting (MOBILE *1*) there.

2.5. *Merging Interpretations*

Partial interpretations may be merged automatically by the inference engine whenever a rule requires a *conjunction* of assertions located in separate nodes. However,

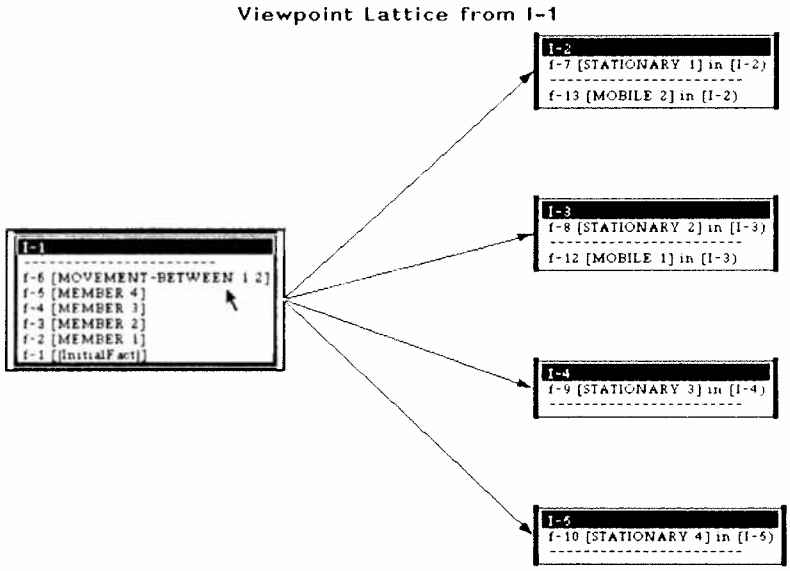

FIG. 6. Local updating of partial hypotheses. Some rule has concluded that if (STATIONARY 1) then (MOBILE 2) must be true and vice versa. Nodes I-2 and I-3 contain these conclusions. Inside the nodes, original hypotheses are shown above a dashed line and the consequences are shown below that line.

merging arbitrary nodes may result in partial interpretations that are inconsistent, e.g., if an entity is labeled as both stationary *and* mobile. Merging nodes I-2 and I-3 in Fig. 6 would create such an inconsistency. The problem is handled by a set of local conflict resolution rules, which detect inconsistent nodes and remove ("poison") them permanently; e.g.,

```
(defrule REMOVE-STATIONARY-AND-MOBILE
    (STATIONARY ?x)
    (MOBILE ?x)
  =>
    (poison)). <remove this node
    permanently>
```

Note that whenever a node is poisoned, all inferior nodes are permanently removed as well. More complex decisions are necessary for conflicts that cannot be resolved locally [19], as we discuss later.

Complete interpretations are assembled on demand by merging all possible combinations of partial interpretations. The following simple rule initiates the necessary merges and marks the resulting complete interpretations with the fact (COMPLETE):

```
(defrule FIND-COMPLETE-INTERPRETATIONS
    (forall (MEMBER ?x)
      (STATIONARY | MOBILE ?x))
  =>
    (assert (COMPLETE))).
```

Figure 7 shows the result of applying this rule to the interpretation graph of Fig. 5. Nodes I-8 and I-10 have been marked as complete interpretations; I-7 and I-9 are intermediate nodes created by the automatic merging process. Inconsistent intermediate nodes that may have been created by this process were removed by local conflict resolution rules.

2.6 *Conflict Resolution*

Local Conflict Resolution. In the previous subsection using a simple example we showed how conflicting *partial* interpretations are removed from the model by executing a constraint rule, which simply *poisons* this particular interpretation. This has been referred to as *local* conflict resolution, because the action is executed inside a particular interpretation regardless of the global state of the model. The content of an interpretation can be seen as a set of *premises* and a (possibly empty) set of *conclusions* that follow from the premises:

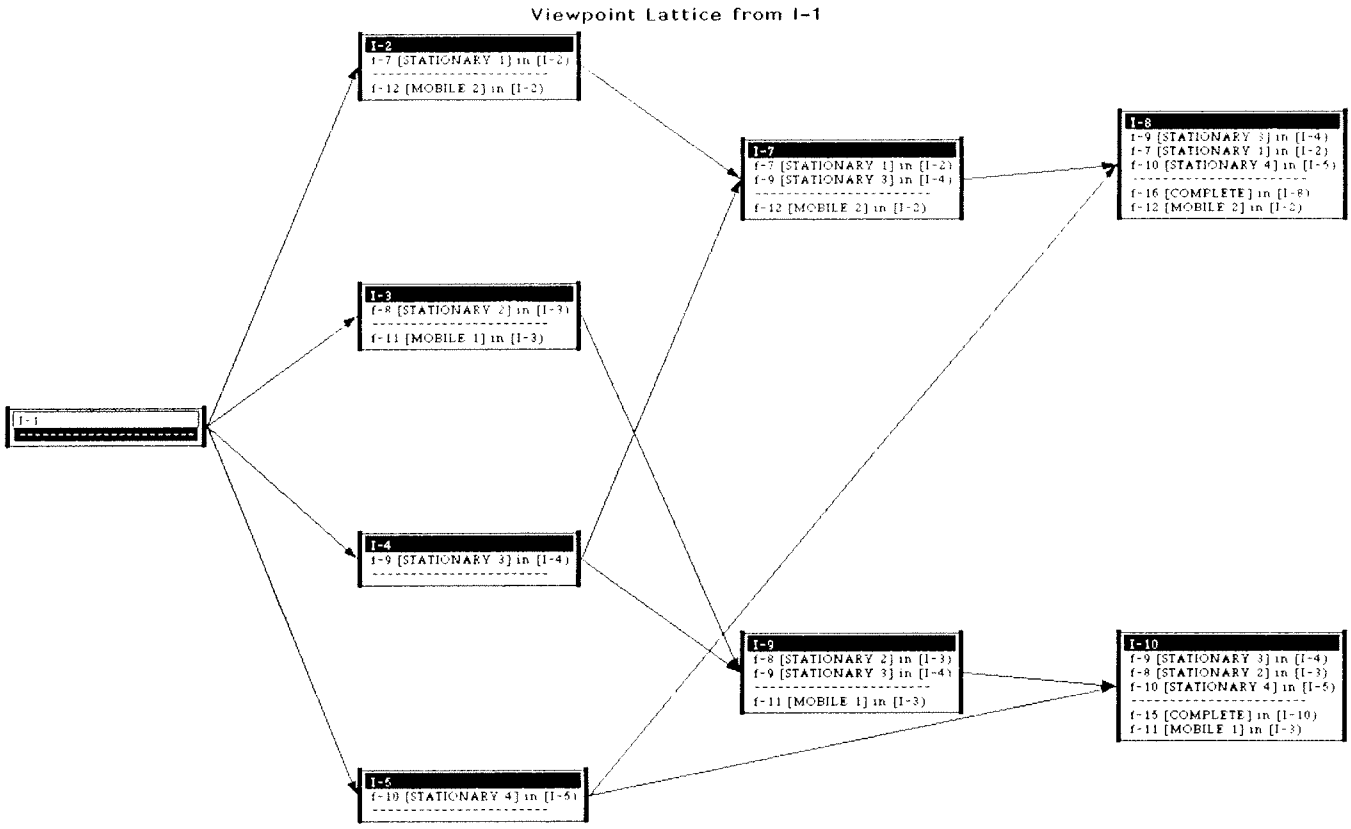

FIG. 7. Finding complete interpretations. Two complete interpretations (nodes I-8 and I-10) have been created by merging partial interpretations. Note that I-2 and I-3 cannot be merged, since these nodes contain mutually conflicting interpretations.

{*premise*-1, *premise*-2, . . .}
→ {*conclusion*-1, *conclusion*-2, . . .}

For instance

{(STATIONARY *1*)} → {(MOBILE *2*)}

in the previous example. A conflict occurs in an interpretation, when the conjunction of premises and consequences can be proven to be false, i.e.,

¬{*premise*-1, *premise*-2, . . . ,
conclusion-1, *conclusion*-2, . . . ,}

as in the case of merging interpretations I-2 and I-3 in Fig. 6:

¬{(STATIONARY 1), (STATIONARY 2),
(MOBILE 1), (MOBILE 2)}.

Assuming that the reasoning step from premises to conclusions is correct, a conflict indicates that the premises (i.e., *hypotheses*) of this particular interpretation cannot be true. Therefore, it is a legitimate move to eliminate this interpretation from the model by poisoning it. In case of a stationary/mobile conflict no other action is required, since the integrity of the whole model is not affected.

Global Conflict Resolution. In some cases the detection of false premises may lead to consequences beyond the interpretation in which the conflict originally occurred. In particular, when premises inherited from previous interpretations are proven false, actions must take place where the false facts were asserted first. This might change the entire structure of the interpretation graph (IG).

The following example demonstrates the problem of global conflict resolution on a scene model containing three features a, b, c. Initially (at $t = t_0$) nothing is known about spatial relationships between these points and whether they are stationary or not. By default they are assumed to be stationary. The initial interpretation of the scene thus contains only

Interpretation $\mathbf{A}(t_0)$:
(STATIONARY a)
(STATIONARY b)
(STATIONARY c).

The states of the model as they develop over time are shown symbolically in Fig. 8. Entities considered stationary are drawn with regular circles while mobile entities are drawn with heavy circles.

Suppose that between t_0 and t_1 all three points show some amount of expansion away from the FOE, giving rise to the conclusion that a is closer (to the vehicle) than b, a is closer than c, and c is closer than b. From the information gathered up to this point, the complete interpretation of the scene at time t_1 looks like this:

Interpretation $\mathbf{A}(t_1)$:
(STATIONARY a)(STATIONARY b)(STATIONARY c)
(CLOSER a b)
(CLOSER a c)
(CLOSER c b)

At time t_2 one of the rules claims that c is closer than a and tries to assert this fact into the current interpretation. Clearly, the new interpretation would contain the conflicting facts (CLOSER a c) and (CLOSER c a), which would not be a feasible interpretation. The conflict is re-

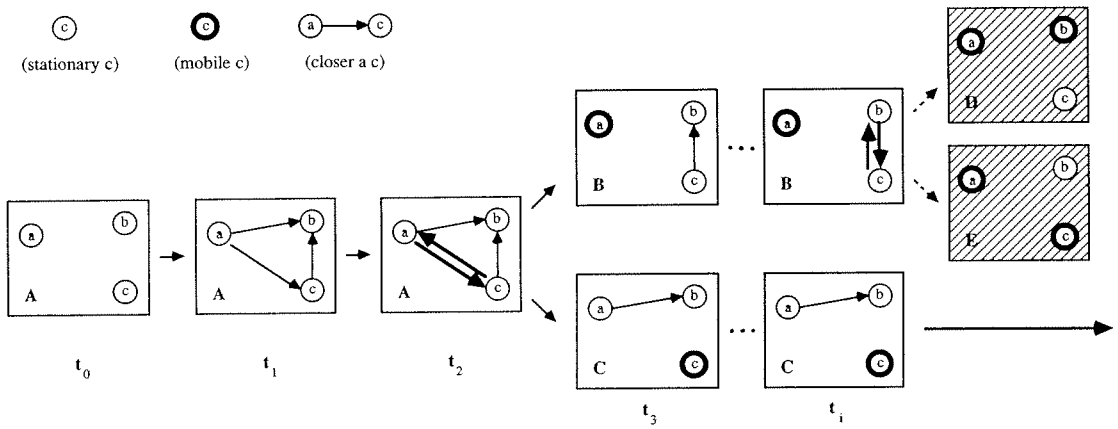

FIG. 8. Development of the Qualitative Scene Model (QSM) over time. At time t_0 three features a, b, c are given, which are initially assumed to be stationary. At time t_1 three CLOSER-relationships have been established between a, b, and c. At time t_2 a conflict occurs in interpretation A by the contradictory facts (CLOSER a c) and (CLOSER c a). Two new interpretations (B and C) are created, each containing one feature considered *mobile* (a, c respectively). At time t_i a new conflict occurs in interpretation B from the additional fact (CLOSER b c). Since another interpretation (C) exists at the same time which could absorb this fact (c is *mobile* in C), B is not branched out but discontinued. C remains the only active interpretation in the model.

solved by creating two disjunct hypotheses **B** and **C**, with either a or c as mobile:

Interpretation **B**(t_2):
(MOBILE a)
(STATIONARY b)
(STATIONARY c)
(CLOSER c b).

Interpretation **C**(t_2):
(MOBILE c)
(STATIONARY a)
(STATIONARY b)
(CLOSER a b).

Since the CLOSER relationship is only meaningful between stationary entities, hypothesizing an entity as being mobile causes the removal of all CLOSER-relationships that exist in the interpretation with respect to this entity.

At this point in time (t_3), *two* feasible interpretations of the scene are active simultaneously. All active interpretations are pursued until they enter a conflicting state, in which case they are either branched into new interpretations or removed from the QSM.

In this example, it is assumed that both interpretation **B** and **C** are still alive at some time t_i. At time t_i, a rule claims that if both b and c are stationary, then b must be closer than c. This would not create a conflict in interpretation **C**, because there c is considered as being MOBILE, making (CLOSER b c) meaningless in this context.

Interpretation **B**, however, cannot ignore this new finding (CLOSER b c) because it considers both **B** and **C** as being stationary and contains the contradictory fact (CLOSER c b)! Again, interpretation **B** could be branched into two new interpretations, with either (MOBILE b) or (MOBILE c). This time, however, there is another active interpretation (**C**), which could absorb (CLOSER b c) without causing an internal conflict. Thus interpretation **B** is not branched out, but removed altogether from the model.

The decision not to create the additional interpretations **D** and **E** of course does not imply that the surviving interpretation **C** is actually correct. As a matter of fact, any of those three interpretations may be the correct one and in general there would be a large number of additional candidates. Any feasible interpretation, i.e., one that cannot be proven false, may also be correct. Since the number of interpretations grows exponentially with the number of entities in the scene, the search for plausible interpretations is subject to the guidelines (metarules) described in Section 2.7.

In the following it is shown how the conflicts in this example would have been actually resolved in terms of partial interpretations. The structure of the interpretation at time t_i is shown in Fig. 9.

The first conflict occurs at time t_2, when (CLOSER c a) is asserted. Since the premises for this new fact are

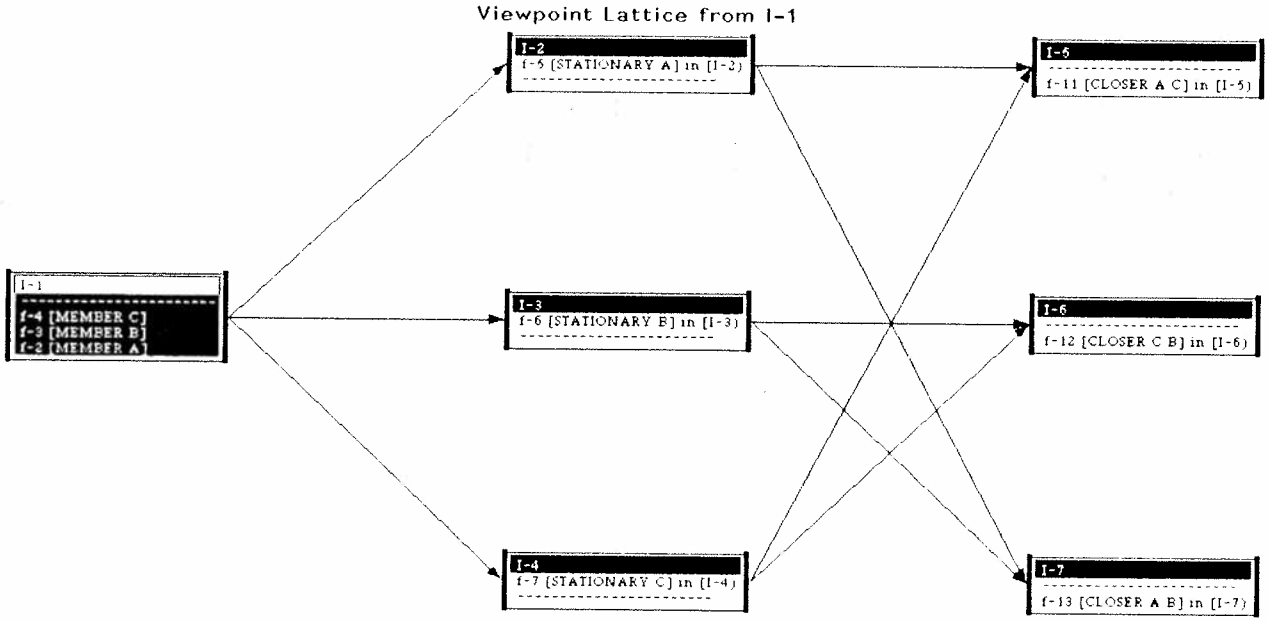

FIG. 9. Structure for the situation at time t_1. CLOSER-relationships are established between three pairs of entities (I-5, I-6, and I-7). Entities a, b, and c are shown as A, B, and C, respectively, in the implemented structure displayed in Figs. 9–12.

(STATIONARY *a*) and (STATIONARY *c*), it is asserted into existing interpretation I-5 which contains exactly those hypotheses. This situation is shown in Fig. 10(a). The conflict suggests that the premise which is the basis of interpretation I-5, namely (STATIONARY *a*) and (STATIONARY *c*), is false. Since *a* and *b* may not be considered both stationary, at least one of them must be moving relative to the other. However, simply removing the conflicting interpretation would not automatically create this conclusion. The problem is solved indirectly by the following rule, which detects the local conflict and spreads the conclusion over the entire model by asserting it in the root node:

```
(defrule CONCLUDE-MOTION-FROM-CLOSER-CON-
FLICT
      (CLOSER ?X ?Y)
      (CLOSER ?Y ?X)
      (at ROOT (assert (MOVEMENT-BETWEEN
      ?X ?Y))))
```

Two new facts are asserted in the root node because of the symmetry in the rule's left-hand side (Fig. 10(b)). In reaction to the new facts in the root, the rules RELATIVE-MOTION-*X* and RELATIVE-MOTION-*Y* conclude (MOBILE *c*) in I-2 and (MOBILE *a*) in I-4 (Fig. 10(c)). This leads to the poisoning of interpretation I-5 in response to its internal stationary/mobile conflict (Fig. 10(d)). Now the original conflict at time t_2 is eventually resolved. The interpretation I-6 and I-7 contain the two separate scene interpretations **B** and **C** respectively.

The second conflict in this example (at time t_i) is caused by the assertion of (CLOSER *b c*) in interpretation I-6 (Fig. 11(a)). As in the previous case, two new facts (MOVEMENT-BETWEEN *b c*) and (MOVEMENT-BETWEEN *c b*) are asserted in the root node, which eventually poisons interpretation I-6 (Fig. 11(b)). Since I-4 labels every member entity of the model and is free of conflicts, it represents a feasible scene interpretation, as does interpretation I-7. In I-7 more entities (2) are stationary than in I-4 (1), such that I-4 might be dropped in favor of I-7 (Fig. 12). The actual implementation, however, would not discard I-4 at this point in time, considering the small evidence in favor of I-7.

In order to allow any intelligent decisions, the information kept in the QSM must be made explicit on demand. In the simplest case, only one scene interpretation exists. When several scene interpretations are feasible at the same time, they must be evaluated according to specific criteria, depending upon the kind of decision that must be made. Often, this does not even require the formation of *complete* interpretations at all.

For example, the FOE computation depends upon a set of image features that are likely to belong to the static environment. Consequently, we can use those features that are not considered *mobile* in any existing interpretation. This set can be found by looking at the partial interpretations only, without building complete interpretations. Similarly, in an alert situation, we may want to know all *potentially moving* entities in the scene, regardless of whether they are all mobile in any complete interpretation. Alternatively, we may rank each complete hypothesis by some "plausibility" measure, e.g., the number of current stationary entities.

2.7. Meta Rules

The process of building the QSM involves four different forms of activities: (a) deriving 3-D facts from the 2-D image sequence, (b) creating hypotheses about the scene, (c) detecting conflicting hypotheses, and (d) resolving those conflicts. In order to avoid a combinatorial explosion of possible scene interpretations, the search for the most plausible scene interpretation is guided by the following *meta rules*:

- Always tend towards the "most stationary" (i.e., most *conservative*) solution. By default, all new entities (i.e., features entering the field of view) are considered stationary.
- Assume that an interpretation is feasible unless it can be proved to be false (the principle of "lack of conflict")
- If a new conclusion causes a conflict in one but not in another current interpretation, then remove the conflicting interpretation.
- If a new conclusion cannot be accommodated by any current interpretation, then create a new, feasible interpretation and remove the conflicting ones.

In this section we described the basic elements, structure, and update mechanisms of the Qualitative Scene Model. In the following section, the knowledge sources are described, which actually create the information stored in the QSM.

3. KNOWLEDGE SOURCES

The QSM serves as the blackboard in a rule-based inference system and is maintained by a generate-and-test process. The two major knowledge sources forming the reasoning engine are the *Generation Rules*, which create partial hypotheses from observations made in the image, and the *Verification Rules*, which try to confirm (or disprove) existing hypotheses (see Fig. 3). In practice, both rule sets are active concurrently and every rule may fire at any time. However, the two categories of rules are distinguished by the way they are formulated.

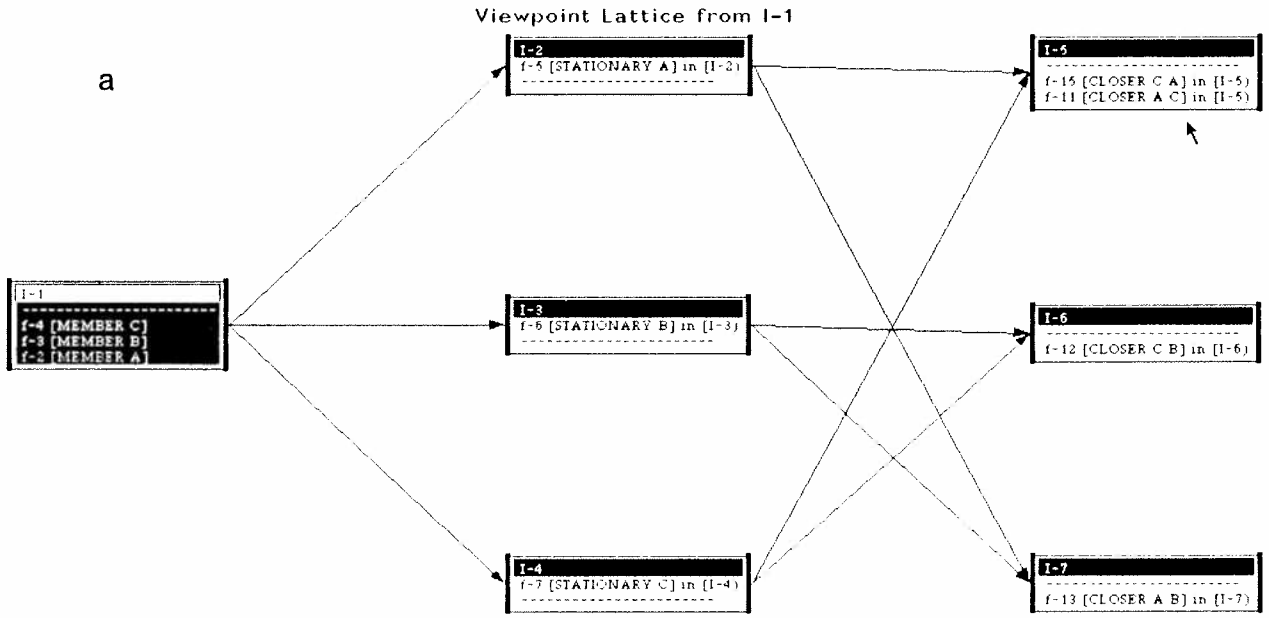

FIG. 10a. Conflict 1. The first conflict occurs at time t_2 when the (CLOSER *c a*) is asserted in I-5 (arrow), which contradicts the existing conclusion (CLOSER *a c*).

Generation Rules are forward chained and try to detect significant events in the input data (Fuzzy FOE and displacement vectors) in a bottom-up fashion. In contrast, the *Verification Rules* are generally backward chained and attempt to collect visual evidence that supports or contradicts an existing hypothesis. Bottom-up rules serve as filters to reduce the search space for possible interpretations, but may produce a large number of irrelevant facts. While a careful balance between bottom-up and top-down execution is a critical design problem, we do not make this classification for the rules discussed below. Instead, we distinguish between rules for making image observations, rules for static scene interpretation, and rules for motion detection and analysis.

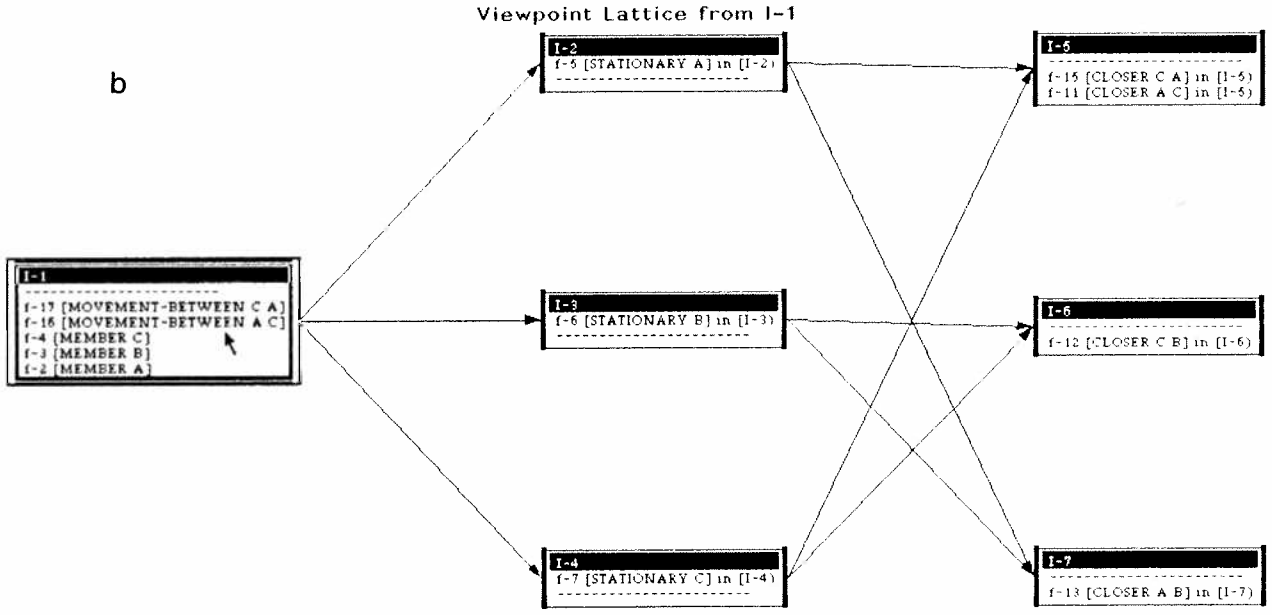

FIG. 10b. Conflict 1. The conflict in I-5 has been detected by the rule CONCLUDE-MOTION-FROM-CLOSER-CONFLICT, which asserted its conclusions as two new facts at the root node I-1 (arrow). The original conflict is not removed yet.

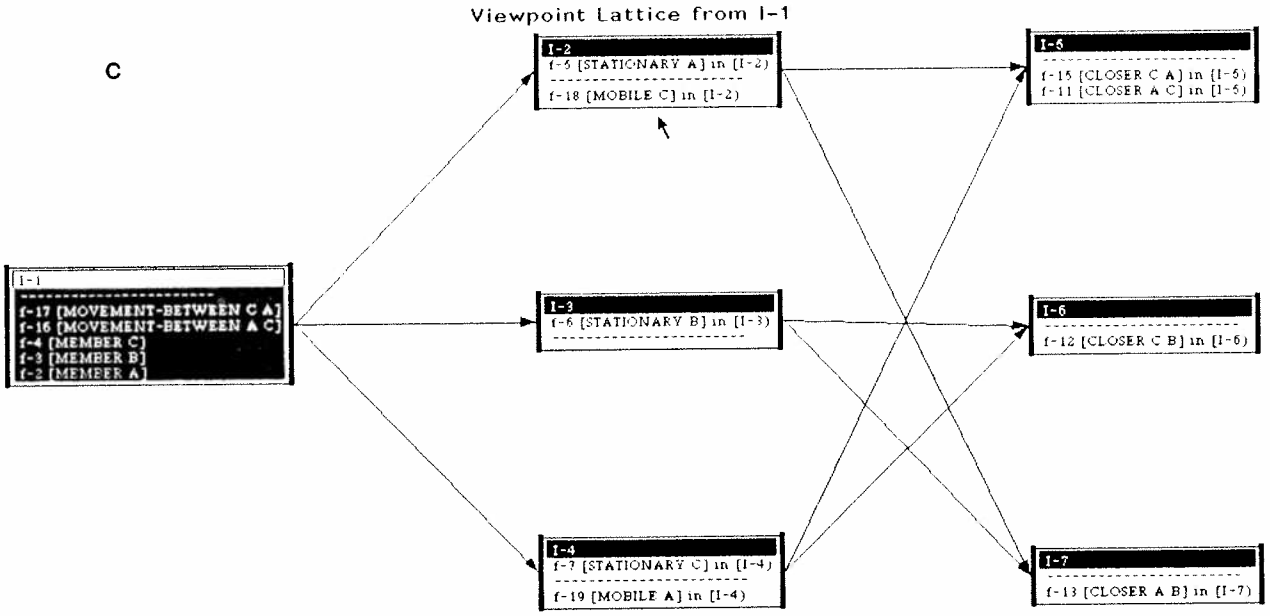

FIG. 10c. Conflict 1. As a consequence of the facts (MOVEMENT-BETWEEN. . .) in the root node, entities *a* and *c* are concluded to be MOBILE in I-4 and I-2 (arrow).

3.1. *Image Observations*

The purpose of the *Image Observation Rules* is to describe in abstract terms certain *static* and *dynamic* 2-D relationships between the features tracked through the image sequence. Static observations are simply derived from the numerical positions of image features, such as between pairs of features. The most important static relations are the following:

Features with respect to each other—

(LEFT-OF *a b t*), (RIGHT-OF *a b t*) feature *a* is left (right) of *b* at time *t*,

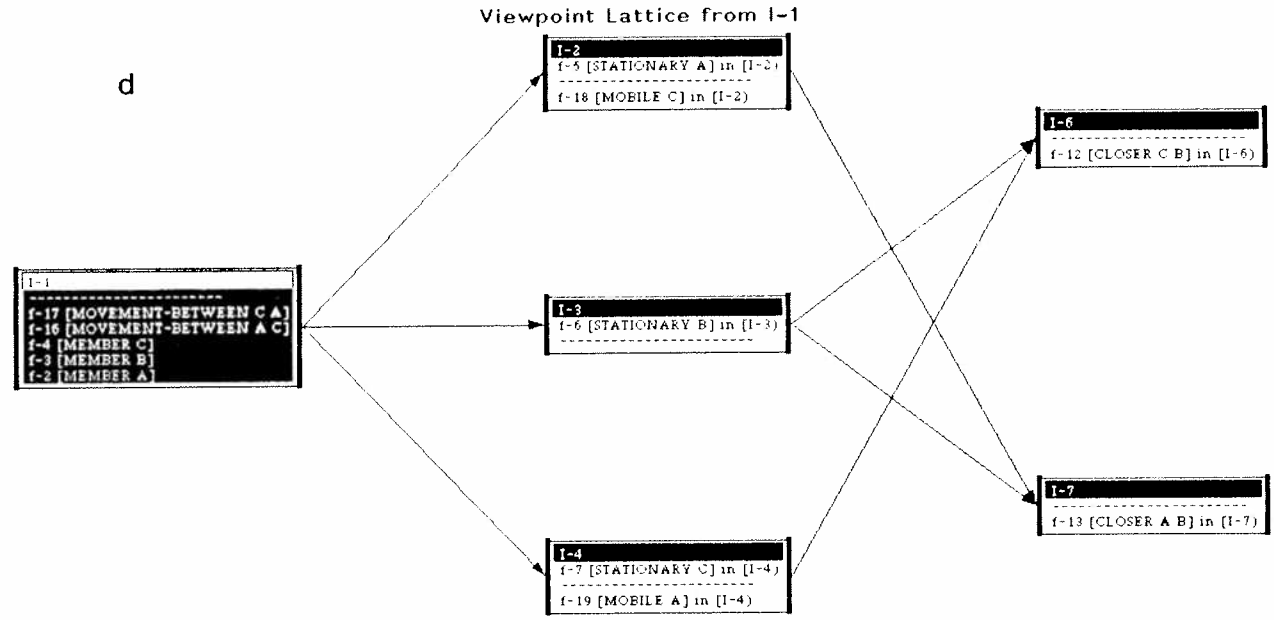

FIG. 10d. Conflict 1. Now the Conflict is finally removed by poisoning I-5. The remaining structures I-6 and I-7 represent the two new scene interpretations B and C.

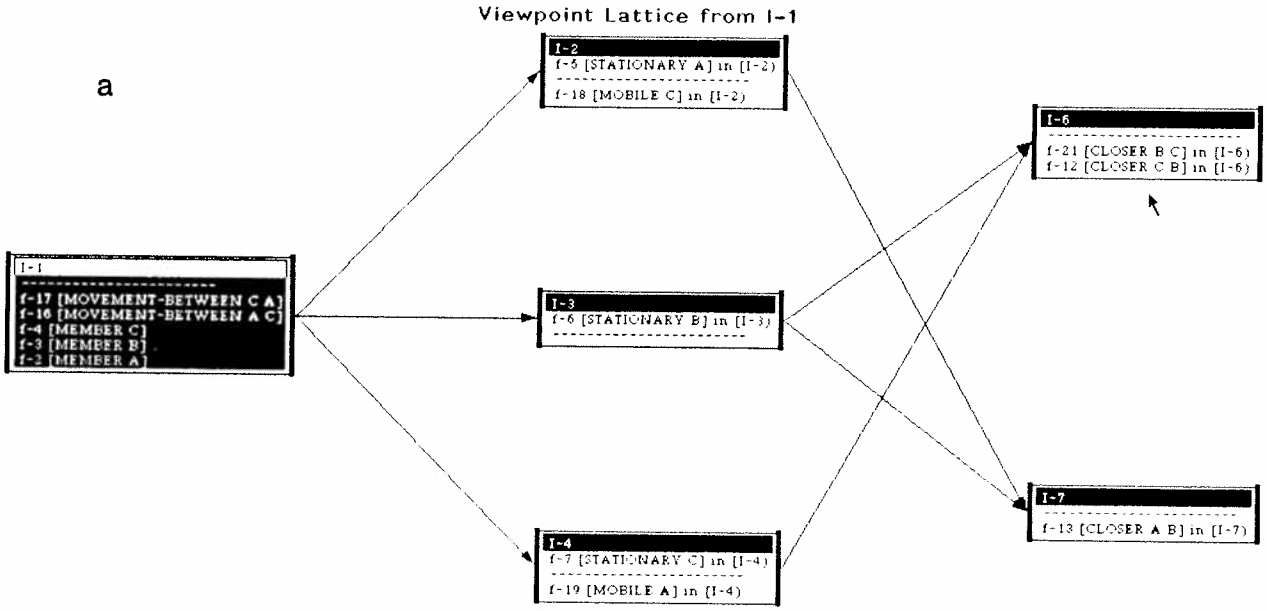

FIG. 11a. Conflict 2. The conflict arises from the conclusion (CLOSER *b c*) in I-6.

(ABOVE *a b t*), (BELOW *a b t*) feature *a* is above (below) *b* at time *t*,

Features with respect to the Fuzzy FOE—

(LEFT-OF-FOE *a t*), (RIGHT-OF-FOE *a t*) feature *a* is left (right) of the FOE at time *t*,

(ABOVE-FOE *a t*), (BELOW-FOE *a t*) feature *a* is above (below) the FOE at time *t*.

Since in general a single FOE-location is not given, the above relationships must be interpreted with respect to a set of possible FOE-locations. For example, (LEFT-OF-FOE *a t*) is true when *a* is left of every possible FOE-location $\mathbf{x}_f$ at time *t*:

(LEFT-OF-FOE *a t*): $x_a < x_f$ for all $\mathbf{x}_f = (x_f y_f) \in \text{FOE}(t)$.

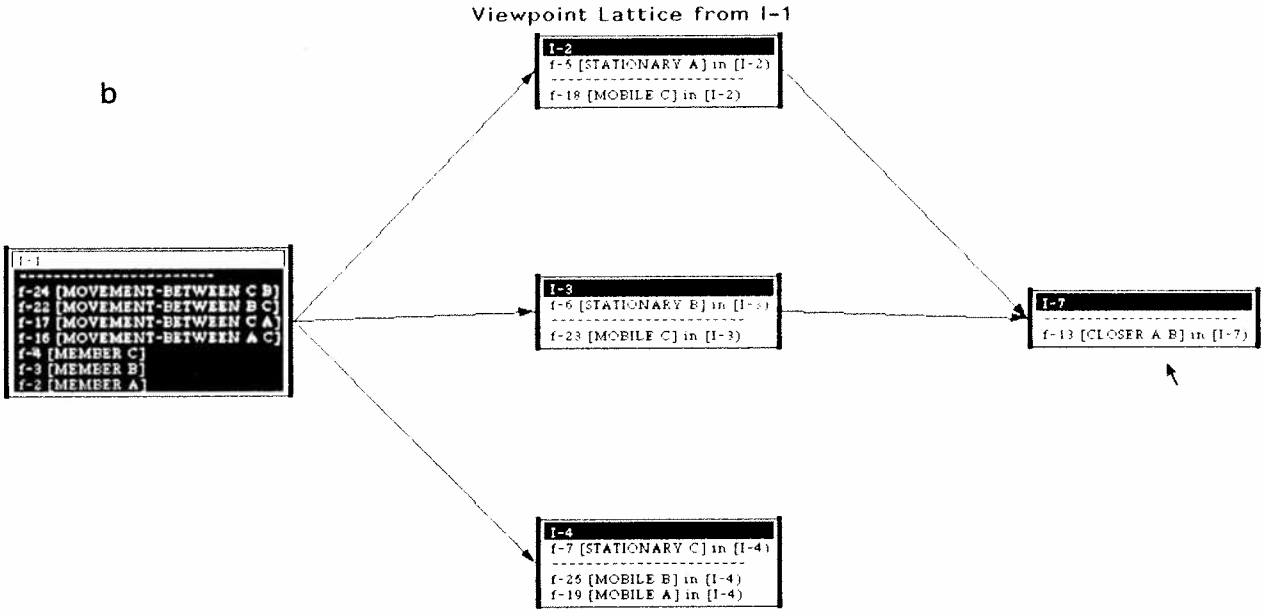

FIG. 11b. Conflict 2. In response to the conflict in I-6, two new facts (MOVEMENT-BETWEEN *b c*) and (MOVEMENT-BETWEEN *c b*) have been asserted at the root node. This, in return, leads to the removal of I-6 and the conclusion of (MOBILE *c*) in I-3 and (MOBILE *b*) in I-4. I-4 and I-7 represent two complete scene interpretations, each providing a label for every entity in the model.

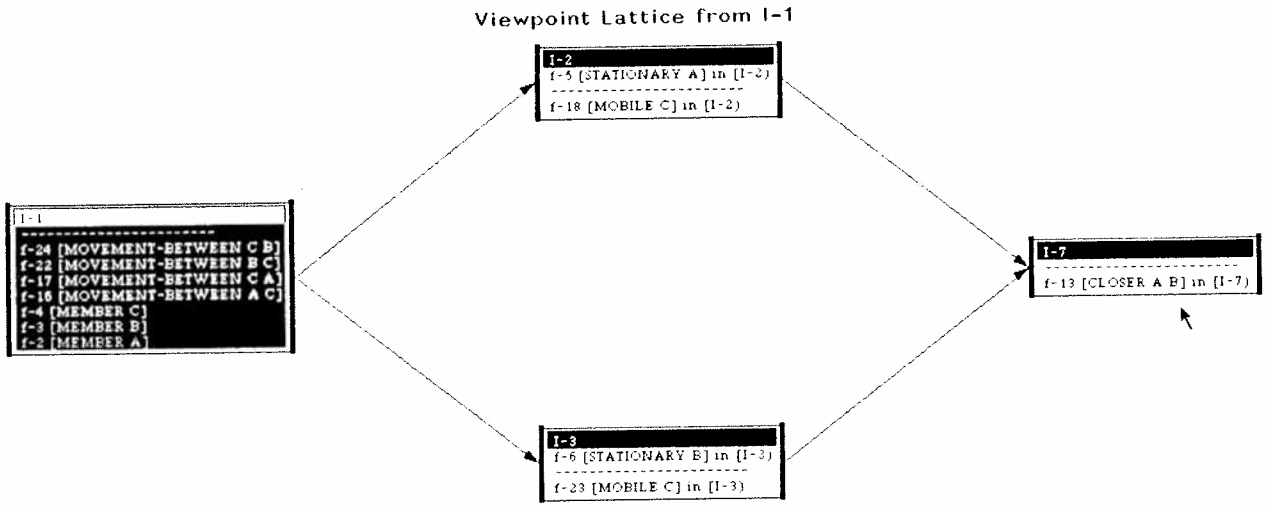

FIG. 12. Interpretation I-4 has been dropped because the interpretation available in I-7 considers more entities (2) as being stationary.

Two other static relationships derived from the ones above are

(OPPOSITE-TO-FOE a b t)—features a and b lie on opposite sides of the FOE, equivalent to (LEFT-OF-FOE a t) and (RIGHT-OF-FOE b t) or to (ABOVE-FOE a t) and (BELOW-FOE b t) (this relation is symmetric, i.e., (OPPOSITE a b t) $\Rightarrow$ (OPPOSITE b a t)),

and

(INSIDE-TO-FOE a b t) features a and b lie on the same side of the FOE but a is closer to the Fuzzy FOE than b.

Again, this is measured relative to the set of possible FOE-locations:

(INSIDE-TO-FOE a b t):

$$\mathbf{d}(\mathbf{x}_f, \mathbf{x}_a) < \mathbf{d}(\mathbf{x}_f, \mathbf{x}_b) \text{ for all } \mathbf{x}_f \in \text{FOE}(t) \leftrightarrow \mathbf{d}(\mathbf{x}_f^{b,\min}, a) < \mathbf{d}(\mathbf{x}_f^{b,\min}, b) \quad \text{and} \quad \mathbf{d}(\mathbf{x}_f^{a,\min}, a) < \mathbf{d}(\mathbf{x}_f^{a,\min}, b),$$

where $\mathbf{x}_f^{b,\min}$ is the FOE-location closest to $\mathbf{x}_b$ ($\mathbf{d}$ is the Euclidean distance in 2-D).

This relationship is particularly easy to determine when the two features are located in a small neighborhood (Fig. 13). For two features lying in different parts of the image, the INSIDE-relationships can only be established when one feature is clearly closer to any possible FOE-location than the other feature. The above formulation takes this into account without explicitly distinguishing the two cases.

Dynamic observations express significant changes that occur in the image between successive frames or over multiple frames. The most important members of this category are the following:

(MOVING-TOWARDS-FOE a t)—feature a moves towards the Fuzzy FOE at time t. This is a strong indicator that the corresponding entity is actually moving in 3-D.

(DIVERGING-NONRADIAL a t)—feature a shows a strong deviation from radial motion, incompatible with a stationary interpretation.

(CONVERGING a b t)—the distance between the two features a and b is getting smaller. This does not imply that either of the two features is actually moving in 3-D. The conclusions drawn from this fact depend upon the context (i.e., particular location of the two features relative to the FOE). This relation is symmetric, i.e.,

(CONVERGING a b t) $\Rightarrow$ (CONVERGING b a t).

(DIVERGING-FASTER a b t)—feature a appears to be moving away from the FUZZY-FOE at a higher rate than feature b. For an exact FOE, we define the diver-

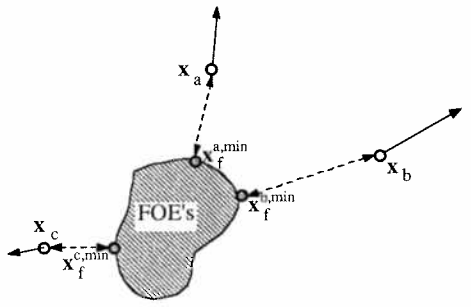

FIG. 13. Establishing the INSIDE-relationship. For image features a, b, c with respect to a given area of possible FOE-locations. $\mathbf{x}_f^{b,\min}$ is the FOE which is closest to b. Here (INSIDE-TO-FOE a b t) is true, but neither (INSIDE-TO-FOE a c t) nor (INSIDE-TO-FOE b c t).

gence $\mathrm{div}(a, t)$ of a feature a at time t as

$$\mathrm{div}(a, t) = \frac{r(a, t) - r(a, t-1)}{r(a, t-1)},$$

where $r(a, t)$ and $r(a, t-1)$ are the 2-D distances between a and the FOE at time t and $t-1$ respectively. With the Fuzzy FOE, the divergence of a feature is expressed as an interval instead of a single value: $\mathrm{div}_{\min}(a, t) \leq \mathrm{div}(a, t) \leq \mathrm{div}_{\max}(a, t)$. If for two features a, b the corresponding intervals do not overlap, then one of them certainly diverges faster than the other, i.e.,

$$\mathrm{div}_{\min}(a, t) > \mathrm{div}_{\max}(b, t) \Rightarrow (\text{DIVERGING-FASTER } a\ b\ t).$$

(PASSING a b t)—feature a is closer to the FOE than b and the two features are getting closer to each other. This is an interesting observation because it supplies strong evidence about the static scene structure. For example, for a person driving a car, close objects (like traffic signs) appear to move toward the periphery of the retina much faster than the background, thus "passing" features at far distance. Fig. 14 shows a typical situation where one feature is passing another in the image.

The following rule for determining the PASSING relationship should demonstrate how some of these different relationships interact:

```
(defrule DETERMINE-PASSES
    (INSIDE-TO-FOE ?A ?B ?t)
    (not (OPPOSITE-TO-FOE ?A ?B ?t))
    (CONVERGING ?A ?B ?t)
  =>
    (assert (PASSING ?A ?B ?t))).
```

Computing any relation involving the FOE must take into account that the Fuzzy FOE stands for a *region* of possible FOE locations, not for a single point. Since these observations must hold for any possible FOE, they may be undetermined for certain features. Consequently, for a small Fuzzy FOE more of these relationships can be deduced than for a large FOE.

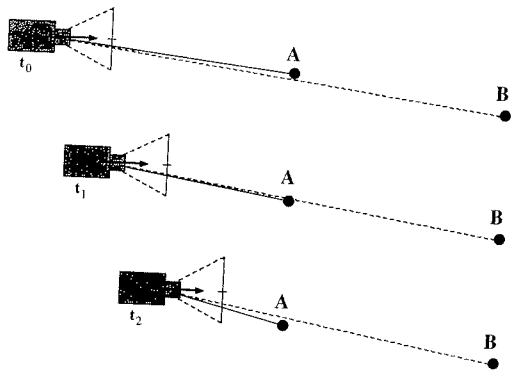

FIG. 14. Passing. A typical situation where one feature, A, seems to be *passing* another, B, in the image.

3.2. *Static Scene Interpretation*

As mentioned in the previous section, the static part of the QSM is built as a partial ordering of entities by their range, based upon the CLOSER relation. The CLOSER relation is transitive, i.e.,

$$(\text{CLOSER } a\ b) \text{ and } (\text{CLOSER } b\ c) \Rightarrow (\text{CLOSER } a\ c).$$

If the *exact* location of the FOE is known, the depth (i.e., its 3-D distance from the camera) of a stationary feature a is proportional to the rate of divergence $\mathrm{div}\ (a, t)$ (see above) of its image [20]. This is the well known basis for *motion stereo*. Obviously, if we know that one feature diverges faster than another, we can conclude that it is also closer to the camera in 3-D, as long as both entities are stationary:

```
(defrule CLOSER-FROM-DIVERGENCE
     (DIVERGING-FASTER ?A ?B ?t) {an image
observation}
    (STATIONARY ?A)
     (STATIONARY ?B) {interpretation with x and
y stationary}
  =>
    (assert (CLOSER ?A ?B))) {a new hypoth-
esis}.
```

While this particular rule is designed to generate hypotheses, a similar rule could be used to verify existing CLOSER hypotheses by checking

$$(\text{CLOSER } a\ b) \Rightarrow \neg(\text{DIVERGING-FASTER } b\ a).$$

Alternatively, CLOSER can be concluded from features "passing" another, which is a special (but obvious) case for different rates of divergence. "Passes" can be detected reliably when features are close to each other in the image, even when the Fuzzy FOE is very large:

```
(defrule CLOSER-FROM-PASSING
    (PASSING ?A ?B ?t)
    (STATIONARY ?A)
    (STATIONARY ?B)
  =>
    (assert (CLOSER ?A ?B))).
```

Again, this rule can be used to verify existing CLOSER hypotheses.

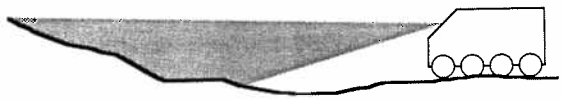

FIG. 15. The convex nature of the viewing profile. It allows the heuristic that features *lower* in the image are generally *closer* to the vehicle.

Supplementary information about the 3-D scene structure can be implanted in a similar fashion, such as hints from occlusion analysis. The following rules make use of the simple heuristic that, with an upright camera in a natural landscape, features *lower* in the image are usually *closer* in 3-D (see Fig. 15). This may, of course, not be valid in other environments, such as indoor scenes:

```
(defrule LOWER-IS-CLOSER
    (BELOW-FOE ?A ?t)
    (BELOW-FOE ?B ?t)
    (BELOW-FOE ?A ?B ?t)
  =>
    (assert (CLOSER ?A ?B))).
```

In the actual implementation, this heuristic rule is used only for the purpose of verifying existing hypotheses. As it turns out in the experiments presented in Section 4, this rule is valuable for detecting implausible static interpretations.

3.3. *Motion Detection and Analysis*

Some forms of 3-D object motion are immediately manifested in the derotated image, whereas other forms of motion require additional reasoning. For example, if an image feature is found to be moving *toward* the Fuzzy FOE (instead of diverging away from it), then it must belong to a moving entity, regardless of its position in 3-D. The corresponding rule contains only one premise and asserts (MOBILE ?x) as a globally known fact (i.e., one that is true in every interpretation):

```
(defrule DIRECT-SINGLE-MOTION-1
    (MOVING-TOWARDS-FOE ?A ?t) {observation
at time t}
  =>
    (assert (MOVES ?A ?t))) {a global fact}.
```

Once an entity has been found *moving*, another rule makes sure that it is remembered as being *mobile* forever:

```
(defrule LABEL-AS-MOBILE
    (MOVES ?A ?t)
  =>
    (assert (MOBILE ?A))) {a global fact}.
```

Since (MOBILE ?A) is asserted globally (i.e., at the root node of the interpretation graph), it is automatically true in every subsequent interpretation. As mentioned in Section 2.5, any partial interpretation containing the conflicting fact (STATIONARY ?A) will automatically be removed by local conflict resolution.

A weaker condition for direct motion detection is given by strongly nonradial image motion of a feature:

```
(defrule DIRECT-SINGLE-MOTION-2
    (DIVERGING-NONRADIAL ?A ?t)
  =>
    (assert (MOVES ?A ?t))).
```

When the Fuzzy FOE is not well defined, the movements of image features relative to the FOE may not be apparent. The rationale for the following rule is that if two entities are static, their images should diverge from the FOE at a rate greater than zero. Consequently, if the FOE is known to lie *between* two features, the features must diverge from each other to permit a static interpretation:

```
(defrule DIRECT-PAIR-MOTION
    (OPPOSITE-TO-FOE ?A ?B ?t)
      (CONVERGING ?A ?B ?t) {if static, they
should diverge}
  =>
    (assert (MOVEMENT-BETWEEN ?A ?B))).
```

The assertion (MOVEMENT-BETWEEN. . .) would in turn fire the rules RELATIVE-MOTION-X and RELATIVE-MOTION-Y to designate one of the features as *mobile* within the appropriate partial interpretations, as described in Section 2.4.

Indirectly, motion can be detected from inconsistent static interpretations. In particular, when there is evidence that one entity (assumed to be static) is CLOSER than another, but there is also evidence for the opposite, a static (i.e., rigid) interpretation is not feasible any longer:

```
(defrule MOTION-FROM-CLOSER-CONFLICT
    (CLOSER ?A ?B) {an inconsistent partial in-
terpretation}
    (CLOSER ?B ?A)
  =>
    (at ROOT (assert (MOVEMENT-BETWEEN ?A
?B)))).
```

The directive "at ROOT" in the above rule causes the subsequent assertion to be placed at the root node of the interpretation graph instead at the node where all the premises are satisfied. Consequently, this form of conflict resolution is *non-local*.

4. EXPERIMENTS

We have implemented a prototype system that runs on a Symbolics 3670 computer. The FOE component was programmed using regular CommonLISP functions. For the rule-based reasoning system, we have used the *ART* development tool [18]. In the following, we demonstrate our approach on two experiments, one with a synthetic image sequence and the other one with real images.

4.1. *Synthetic Example*

The purpose of the first example is to discuss a variety of different situations on a single (synthetic) image sequence, particularly to show how the QSM develops over time. It does not include the results of the FOE computation. The generated image sequence shows a road scene that contains a set of stationary and moving objects (Fig. 16). The camera is moving at 15 km/h towards an intersection, which is initially 80 m away. Frames are taken at 0.5 s intervals, resulting in a forward motion of about 2 m per frame. The camera also performs horizontal and vertical rotations in the range of ±5°.

The scene contains two moving objects. A van (labeled **F**) is approaching the camera on the same road at a velocity of roughly 22 km/h. Another car (labeled **P**) is crossing the path of the camera from right to left at 36 km/h. The initial distance to the static pole **M** is 17 m; the hills in the background are about 600 m away.

The development of the QSM for this image sequence is shown graphically in Fig. 17. For four points in time (0.5, 1.0, 2.0, and 4.5 s), the state of the QSM as actually produced by the reasoning engine is displayed as the set of *complete* interpretations that existed at these moments. Concurrent interpretations are stacked vertically in this figure, but they have not been ranked. *Stationary* entities are labeled with a circle; all other entities are considered *mobile*. Established CLOSER relationships between stationary entities are indicated by a connecting line between the two image features, where the closer entity carries a larger circle. In this example, only the observed "passes" between features were used to con-

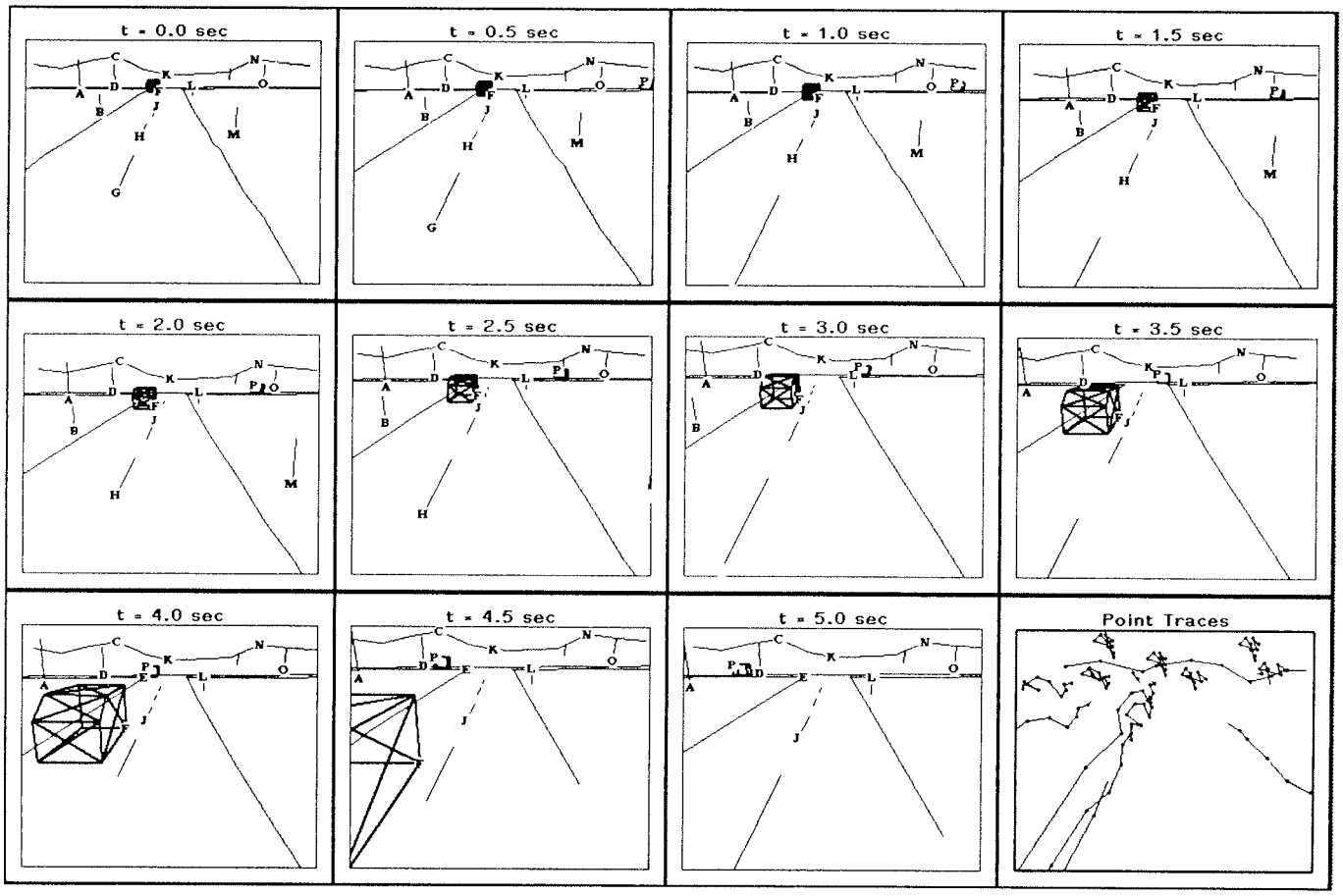

FIG. 16. Synthetic image sequence. The camera is moving towards an intersection. The scene contains two moving objects: a van (marked **F**) which is approaching the camera on the same road, and another car (**P**) which is crossing the path of the camera from right to left. The traces of the point features over the entire sequence are shown in the lower righthand corner.

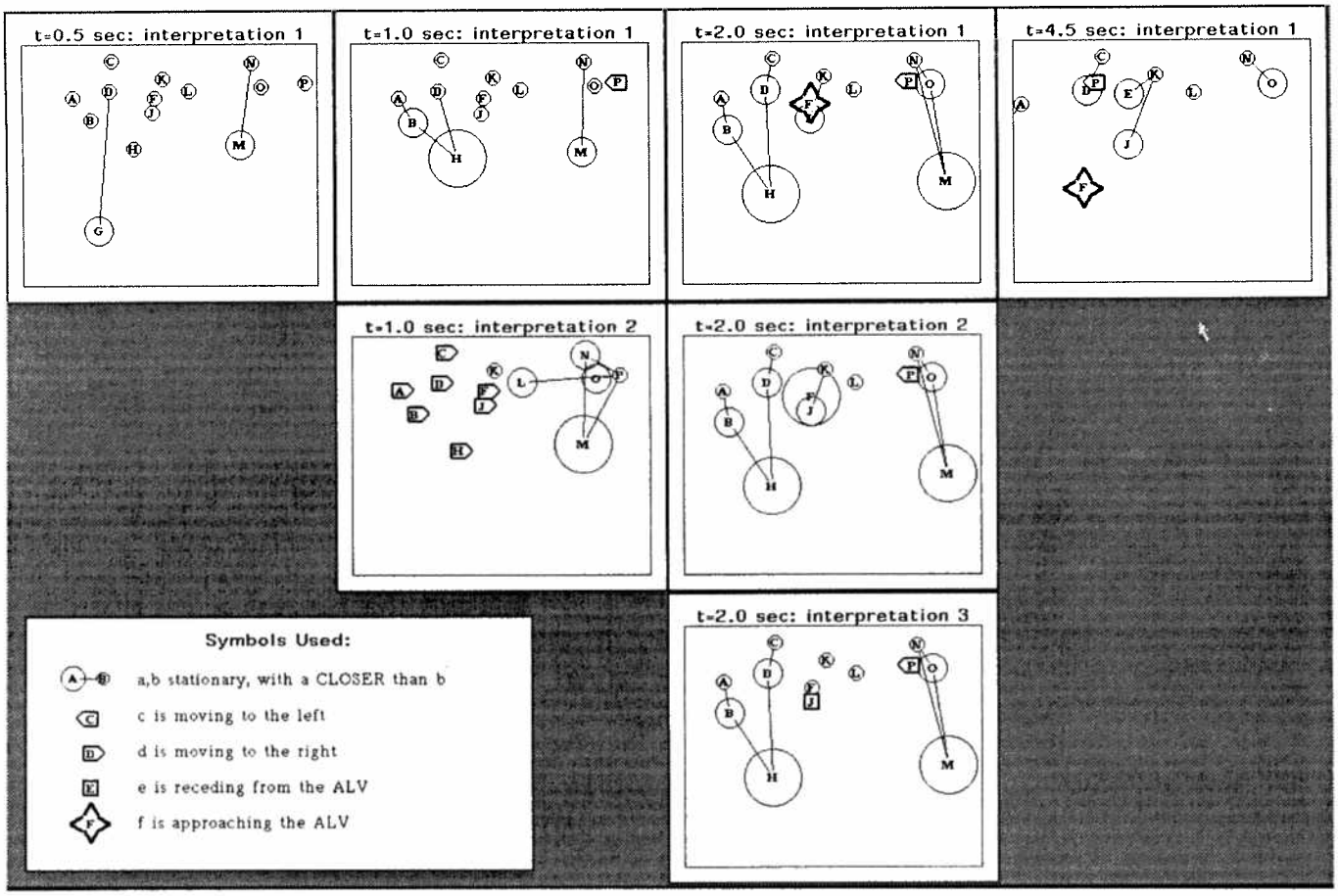

FIG. 17. Interpretation of synthetic image sequence. The complete interpretations contained in the QSM are shown at four points in time (0.5, 1.0, 2.0, 4.5 seconds). Stationary entities are marked with circles. Established closer-relationships are indicated by connecting lines between stationary entities, with the closer entity denoted by the larger circle. Concurrent interpretations are stacked vertically without ranking.

clude CLOSER relationships, not the relative rate of divergence from the FOE. As a result, no CLOSER relationships are found across the image center.

At time $t = 0.5$ s (after the first frame pair), the scene is considered completely static with two CLOSER relationships established.

At time $t = 1.0$ s, the moving car (**P**) enters the field of view. Relative movement between **P** and other features across the FOE leads to the creation of two scene interpretations:

Interpretation 1 considers **P** as mobile and moving to the left.

Interpretation 2 "thinks" **P** is stationary and the other entities (**A**, **B**, **C**, . . .) on the opposite side of the FOE are mobile and moving to the right. The stationary entity **P** has been linked to other entities by CLOSER relationships.

The second interpretation is eliminated after the subsequent frame, when one of the features on the left (**B**) is found to be definitely moving away from the FOE. This movement contradicts the hypothesis that **B** is moving to the right. The approaching van (**F**) has not been detected up to this point. Since it is moving toward the camera approximately on a straight path with constant velocity, its motion is not immediately found.

At time $t = 2.0$ s the motion of the van causes feature **F** (the van) to "pass" feature **J**, which temporarily creates three interpretations:

Interpretation 2 simply says that both **F** and **J** are stationary and that entity **F** is closer than **J**, due to the observed "pass." This interpretation is geometrically feasible.

However, point **J** is lower in the image than **F**, and should therefore be closer than **F** (according to the heuristic LOWER-IS-CLOSER rule). Consequently, motion between the entities **F** and **J** is hypothesized:

Interpretation 1 sees **F** approaching and **J** stationary (the correct solution), while

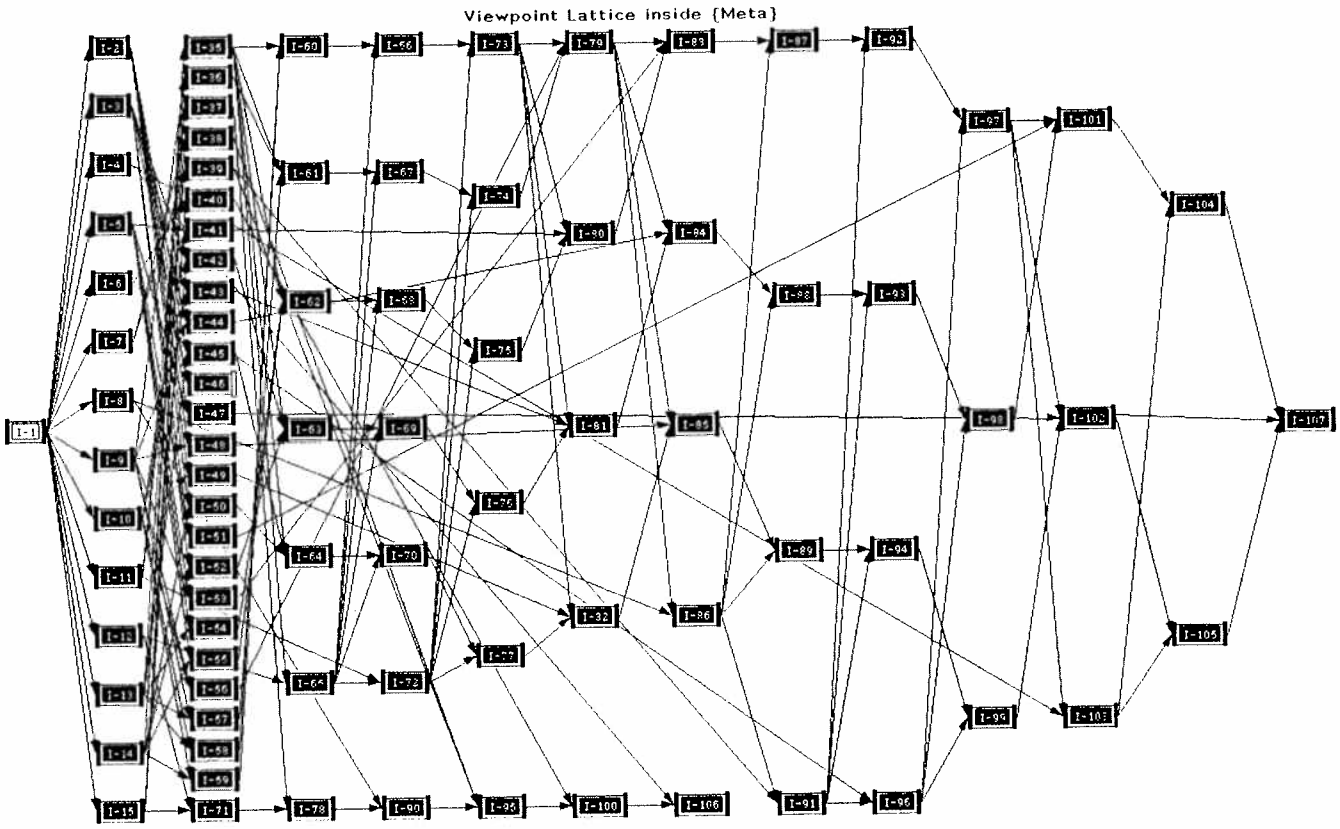

FIG. 18. Part of the interpretation graph for the synthetic example (t = 0.5 sec). I-1 (on the left) is the root node and I-2 to I-15 are the initial default hypotheses for individual entities in the scene. Arrows indicate inheritance between nodes. Not all existing links are shown in this diagram. I-107 (on the right) represents one complete interpretation formed by merging partial interpretations.

Interpretation 3 sees **F** stationary and **J** receding from the camera.

The last interpretation is refuted subsequently, when **J** is found to be diverging downwards from the FOE, which contradicts its receding motion. Eventually, at t = 4.5 s, only one interpretation has survived, showing **F** and **P** as mobile and all other entities as stationary. Fig. 18 shows the interpretation graph for this example at time t = 0.5 s after creating the only existing complete interpretation. Inner nodes of the graph correspond to partial interpretations which are combined to form *complete* interpretation (I-107).

4.2. *Real Example*

For the second example (Fig. 19), we have used an image sequence taken from the *Autonomous Land Vehicle* (ALV) driving on a road through a test site (Fig. 19a). To obtain the original displacement vectors, point features were selected and tracked manually between successive frames. This was done on binary edge images (Fig. 19b) to imitate the conditions for automatic point tracking, because some clues obvious (to humans) in the original grey-scale sequence are lost during edge detection. Consequently, the end points of the original displacement vectors are not very accurate. Recent experiments on extended sequences [21, 22] show that similar displacement vectors can be achieved with fully automatic feature tracking.

Fig. 19c shows the results of the FOE computation. The shaded area near the image center represents the Fuzzy FOE, with the small circle in the middle that marks the location of minimum error FOE point. The feature points used to compute the FOE are those that are *not* considered mobile by any interpretation in the current scene model. The computed rotations for this frame pair (about 0.2° horizontally and 0.1° vertically) are shown in a coordinate square in the lower left-hand corner. The distance traveled over the (assumed planar) road surface between frames has been estimated as 2.1 m.

The scene contains two moving objects. Point 24 belongs to a car which had passed the ALV earlier in this sequence. It is clearly identified as being *mobile* and is

a

FIG. 19a. Real example. Image sequence contains two moving objects: a car that has passed the ALV and is barely visible in the distance, and a second car that is approaching in the opposite direction and is about to pass.

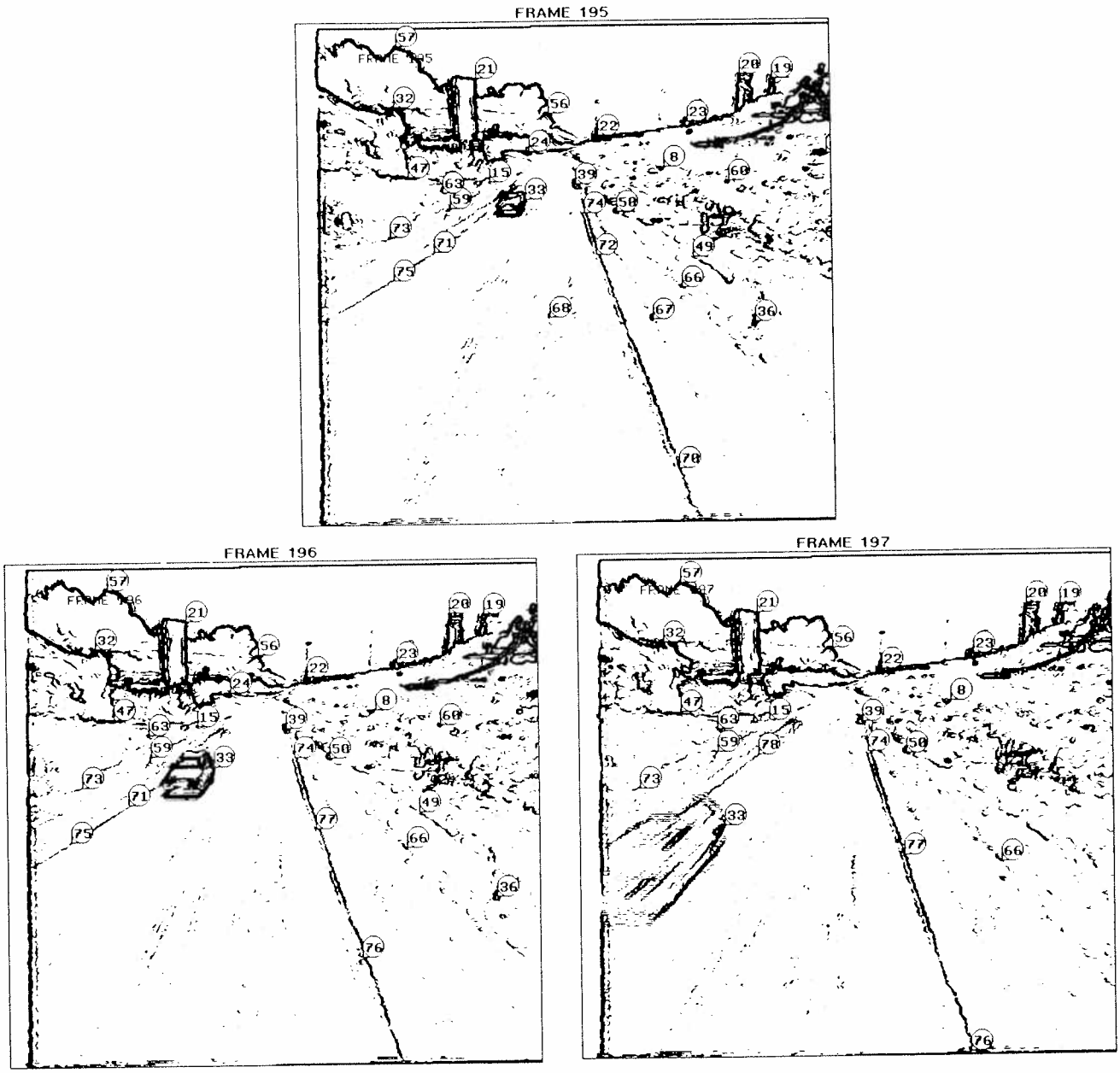

FIG. 19b. Images after edge enhancement and feature point selection. The numbered circles indicate points that are being tracked from image to image. Note that points 24 and 33 are on the moving cars.

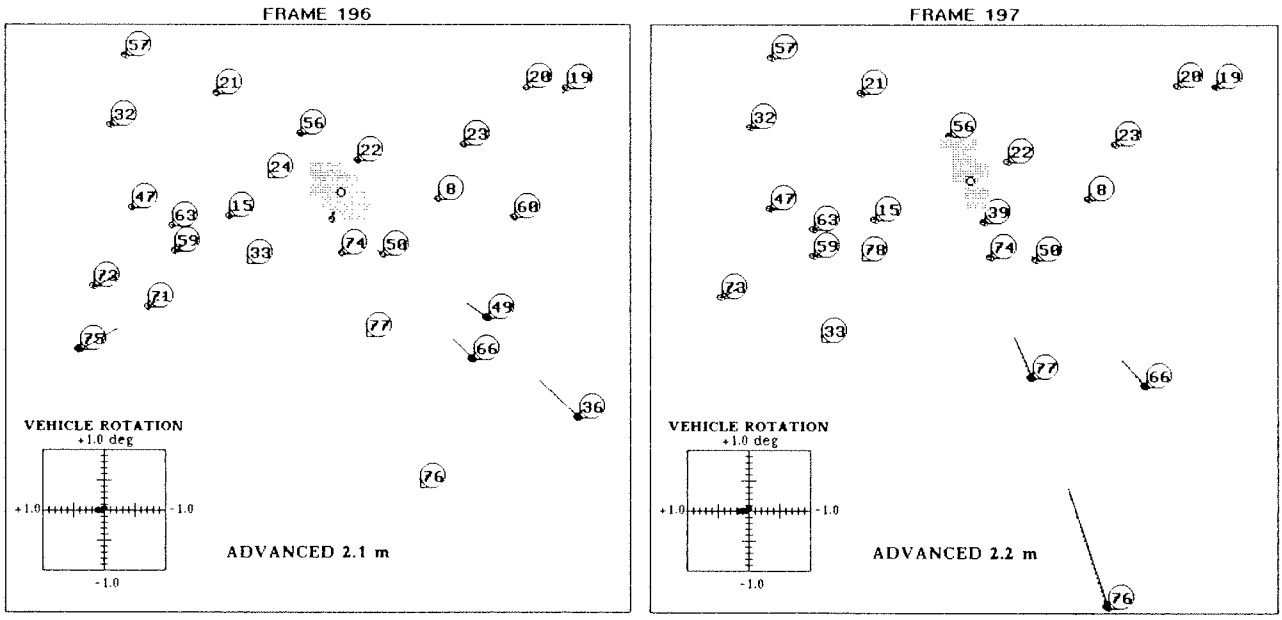

FIG. 19c. Displacement vectors and resulting Fuzzy FOE (shaded area), camera rotations (about two axes) and estimated advancement. The circle inside the shaded area is the estimated FOE location with the lowest error value. Rotation about the third axis is small enough to be neglected.

about to disappear in the current frame. Point 33 is located on a car approaching the ALV, but its motion has not been detected up to the current frame.

At frame 196, some movement between feature 33 and several other features (15, 39, 50, . . . , 73) has been detected but the direction of motion can not be resolved. Two different interpretations are created (Fig. 19d), one with entity 33 as *mobile* (Interpretation 1) and the other with entity 33 as *stationary* (Interpretation 2). Both interpretations are carried over to frame 197 (Fig. 19d), where two significant things happen.

In Interpretation 1 for frame 197 (Fig. 19d), entity 33 is concluded to be approaching the camera, because of its relative position to stationary entities and its downward movement. Thus, Interpretation 1 says that "*if* 33 *is mobile, then it is approaching our vehicle.*"

In Interpretation 2 for frame 197 (Fig. 19d), entity 33 is still regarded as stationary. If this were true, however, then 33 must be quite close to the vehicle, even *closer* than entity 76 (at the bottom of the image)! This situation would be very unlikely (LOWER-IS-CLOSER heuristic) and therefore, Interpretation 2 is ruled out. Only the correct interpretation 1 for frame 197 (Fig. 19d) remains.

5. CONCLUSIONS

The difficulty of understanding dynamic scenes from a moving camera is that *stationary* objects are generally not still in the image while *mobile* objects do not necessarily appear to be in motion. Consequently, the detection of 3-D object motion sometimes requires reasoning far beyond simple 2-D change analysis. In this paper, we presented the conceptual outline of a new approach to scene understanding in dynamic environments.

Our approach departs from related work by following a strategy of *qualitative* rather than quantitative reasoning and modeling. While quantitative techniques have traditionally been dominant in computer vision, qualitative techniques are now receiving growing attention in this field [23, 24]. They hold the potential to replace expensive numerical computations and models by simpler reasoning about the important properties of the scene, representations. This is particularly true for the higher levels of vision and it seems to be a useful methodology for building abstract descriptions gradually, starting at the lowest level.

The numerical effort in our qualitative approach is packed into the computation of the Fuzzy Focus of Expansion (FOE) a low-level process, which is performed entirely in 2-D. We have extended the FOE concept by computing a connected *region* of possible FOE-locations (called the *Fuzzy FOE*), instead of a single point FOE image location. The subsequent reasoning process evaluates the "derotated" displacement field with respect to the Fuzzy FOE and creates a qualitative description of the scene. *Multiple* scene interpretations are pursued concurrently to reflect the ambiguities inherent in any type of scene analysis. If only one interpretation was

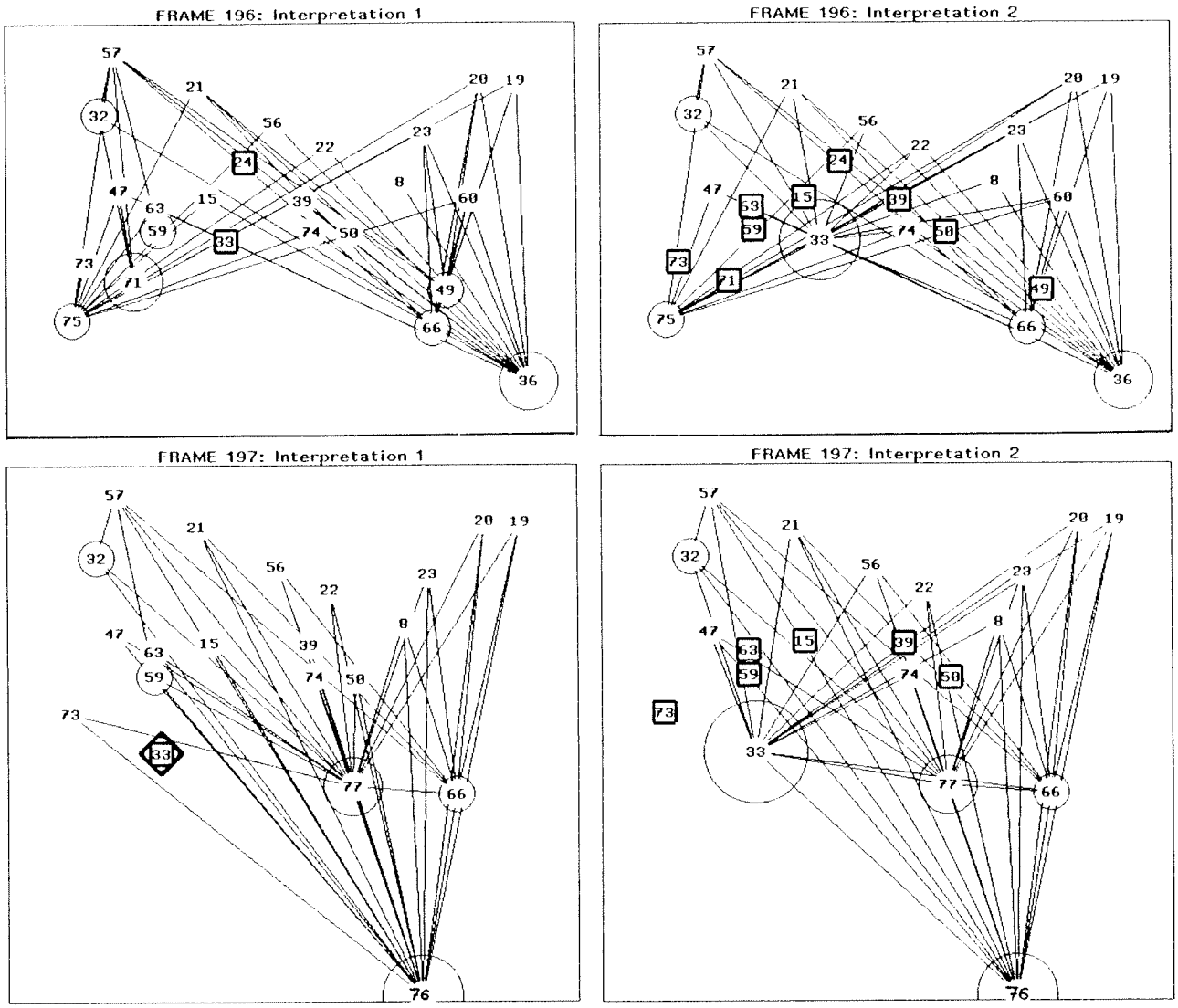

FIG. 19d. Interpretation of real example. Two different scene interpretations for FRAME 196 are created. Entity 24 is known to be moving (from earlier conclusions) in both interpretations, but its direction of motion is currently undetermined (indicated by a square). Interpretation 1 for FRAME 196: entity 33 (square) is considered mobile with undetermined motion. Interpretation 2 for FRAME 196: entities 15, 39, 50, . . . , 73 (squares) are mobile, 33 is stationary. Neither of these interpretations can currently be ruled out and both are carried over to the next frame pair. Interpretation 1 for FRAME 197: entity 33 is concluded to be moving towards the camera (indicated by an upright square). Interpretation 2 for FRAME 197: It is about to vanish. If entity 33 was really stationary, then it must be *closer* to the camera than entity 76 (at the bottom), indicated by the arc from 33 to 76 and the larger circle around 33. However, this contradicts the heuristic that entities lower in the image are generally closer in 3-D space, which makes the entire interpretation implausible.

available at any time, the chance of that interpretation being incorrect would be significant. Simultaneously evaluating a set of scene interpretations allows us to consider several alternatives and, depending upon the situation, select the appropriate one (e.g., the most "plausible" or the most "threatening" interpretation).

For our implementation, we have used off-the-shelf expert system tools mainly because they allow easy manipulation of declarative knowledge. Execution speed was only of minor importance. The examples presented here show the basic operation of this system and demonstrate that some apparently simple situations may actually require complex paths of reasoning.

The availability of reliable displacement vectors is important to our approach. While we used manual point tracking for the examples shown here, recent experiments indicate that *automatic* feature selection and tracking have become practical. The system described here has been successfully applied to ALV image sequences with over 250 frames in a fully automatic mode

[21]. Extending this approach to more complex image features, such as line segments and region boundaries, is a future objective. However, the current focus of our work aims at the integration of multiple sources of information into a reliable and robust framework for dynamic scene analysis [21].

REFERENCES

1. A. R. Bruss, and B. K. P. Horn, Passive navigation, *Comput. Vision Graphics Image Process.* **12,** 1983, 3–20.
2. O. D. Faugeras, F. Lustman, and G. Toscani, Motion and structure from point and line matches, in *Proceedings, 1st International Conference on Computer Vision, ICCV'87, IEEE Computer Society, London, June 1987,* pp. 25–34.
3. H. C. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, *Nature* **293,** 1981, 133–135.
4. A. Mitiche, S. Seida, and J. K. Aggarwal, Determining position and displacement in space from images, in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition CVPR'85, San Francisco, June 1985,* pp. 504–509.
5. R. Y. Tsai and T. S. Huang, Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, *IEEE Trans. Pattern Anal. Mach. Intelligence* **PAMI-6**(1), 1984, 13–27.
6. S. Bharwani, E. Riseman, and A. Hanson, "Refinement of environmental depth maps over multiple frames, in *Proceedings, IEEE Workshop on Motion: Representation and Analysis, May 1986,* pp. 73–80.
7. S. Ullman, *Maximizing Rigidity: The Incremental Recovery of 3-D Structure from Rigid and Rubbery Motion,* A.I. Memo No. 721, MIT Artificial Intelligence Lab, June 1983.
8. G. Adiv, Determining three-dimensional motion and structure from optical flow generated by several moving objects, *IEEE Trans. Pattern Anal. Mach. Intelligence,* **PAMI-7**(4), 1985, 384–401.
9. C. Jerian and R. Jain, Determining motion parameters for scenes with translation and rotation, *IEEE Trans. Pattern Anal. Mach. Intelligence* **PAMI-6**(4), 1984, 523–530.
10. H. C. Longuet-Higgins and K. Prazdny, The interpretation of a moving retinal image, *Proc. Ro. Soc. London B* **208,** 1980, 385–397.
11. K. Prazdny, Determining the instantaneous direction of motion from optical flow generated by a curvilinear moving observer, *Comput. Vision Graphics Image Process.* **17,** 1981, 238–259.
12. R. C. Bolles and H. H. Baker, Epipolar-plane analysis: A technique for analyzing motion sequences, in *Proceedings, IEEE Workshop on Motion: Representation and Control, October 1985,* pp. 168–178.
13. R. Jain, Direct computation of the focus of expansion, *IEEE Trans. Pattern Anal. Mach. Intelligence* **PAMI-5**(1), 1983, 58–64.
14. D. T. Lawton, Processing translational motion sequences, *Comput. Vision Graphics Image Process.* **22,** 1983, 114–116.
15. D. H. Marimont, Projective duality and the analysis of image sequences, in *Proceedings, IEEE Workshop on Motion: Representation and Analysis, May 1986,* pp. 7–14.
16. W. Burger and B. Bhanu, On computing a 'fuzzy' focus of expansion for autonomous navigation, in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, CVPR'89,* pp. 563–568.
17. J. H. Rieger and D. T. Lawton, Processing differential image motion, *J. Opt. Soc. Am. A* **2**(2), 1985, 354–360.
18. B. D. Clayton, *ART Programming Manual,* Inference Corp., Los Angeles, California, 1985.
19. W. Burger and B. Bhanu, Qualitative motion understanding, in *Proceedings 10th International Joint Conference on Artificial Intelligence, IJCAI'87,* pp. 819–821, Morgan Kaufman.
20. K. Prazdny, On the information in optical flows, *Comput. Vision Graphics Image Process.* **22,** 1983, 239–259.
21. B. Bhanu, P. Symosek, J. Ming, W. Burger, H. Nasr, and J. Kim, Qualitative target motion detection and tracking, in *Proceedings, DARPA Image Understanding Workshop, Palo Alto, CA, May 1989,* pp. 370–397, Morgan Kaufman.
22. J. Kim and B. Bhanu, *Motion Disparity Analysis Using Adaptive Windows,* Technical Report 87SRC38, Honeywell Systems & Research Center, Minneapolis, MN, 1987.
23. W. B. Thompson and J. K. Kearney, Inexact vision, in *Proceedings, IEEE Workshop on Motion: Representation and Analysis,* pp. 15–21. May 1986.
24. A. Verri and T. Poggio, Against quantitative optical flow, in *Proceedings, First International Conference on Computer Vision, ICCV'87, London,* pp. 171–180, IEEE Computer Society.