Gewali, L., Meng, A., Mitchell, J., and Ntafos, S. 1988 (Urbana, Ill., June). Path planning in 0/1/infinity weighted regions with applications. *Proceedings of the Fourth Annual ACM Symposium on Computational Geometry.* New York: ACM, pp. 266–278.

Lozano-Pérez, T., and Wesley, M. A. 1979. An algorithm for planning collision-free paths among polyhedral obstacles. *Comm. ACM* 22(10):560–570.

Mitchell, J. S. B. 1987 (New Orleans, La., May). Shortest paths among obstacles, zero-cost regions, and roads. Paper delivered at the Joint National Meeting of TIMS/ORSA.

Mitchell, J. S. B. 1988. An algorithmic approach to some problems in terrain navigation. *Art. Intell.* 37:171–201.

Mitchell, J. S. B., and Keirsey, D. 1984. Planning strategic paths through variable-terrain data. *SPIE Conference on Applications of Artificial Intelligence.* Vol. 485. New York: SPIE, pp. 172–179.

Richbourg, R., Rowe, N., Zyda, M., and McGhee, R. 1987 (Raleigh, N.C., March). Solving global two-dimensional routing problems using Snell's Law and A* search. *Proceedings of the IEEE International Conference on Robotics and Automation.* New York: IEEE, pp. 1631–1636.

Rowe, N. C., and Richbourg, R. F. 1990. An efficient Snell's law method for optimal-path planning across multiple two-dimensional, irregular, homogeneous-cost regions. *Int. J. Robot. Res.* 9(6):48–66.

Qualitative Understanding of Scene Dynamics for Mobile Robots

Wilhelm Burger
Bir Bhanu

Honeywell Systems and Research Center
Minneapolis, Minnesota 55418

Abstract

In this paper, we present a new approach to analyzing motion sequences as they are observed from a mobile robot operating in a dynamic environment. In particular, we address the problems of (a) estimating the robot's egomotion, (b) reconstructing the 3D scene structure, and (c) evaluating the motion of individual objects from a sequence of monocular images. Our approach consists of a two-stage process starting from given sets of displacement vectors between distinct image features in successive frames. First, the robot's egomotion is computed in terms of rotations and the direction of translation. To cope with the problems of noise, we have extended the concept of the Focus of Expansion (FOE) by computing a Fuzzy FOE, which defines an image region rather than a single point. In the second stage, a 3D scene model is constructed by analyzing the movements and positions of image features relative to each other and relative to the Fuzzy FOE. Using a mainly qualitative strategy of reasoning and modeling, multiple scene interpretations are pursued simultaneously. This second stage allows the determination of moving objects in the scene. Results of this approach applied to a real image sequence are presented.

1. Introduction

Visual information plays a key role in mobile robot operation. Even with the use of sophisticated inertial navigation systems, the accumulation of position

This research was supported by DARPA under contract DACA76-86-C0017 and monitored by the U.S. Army Engineer Topographic Laboratories.

errors requires periodic corrections. Operation in unknown environments or mission tasks involving search, rescue, or manipulation critically depend on visual feedback. Motion understanding becomes vital as soon as moving objects are encountered in some form (e.g., while following a convoy, approaching other vehicles, or detecting moving threats). In the given case of a moving camera, image motion can also supply important information about the spatial layout of the environment and the actual movements of the autonomous mobile robot.

Previous work in motion understanding has focused mainly on numeric approaches for the reconstruction of 3D motion and scene structure from 2D image sequences [see Nagel (1986) for a review]. In the classic numeric approach, structure and motion of a rigid object are computed simultaneously from successive perspective views by solving systems of linear or nonlinear equations (Bruss and Horn 1983; Faugeras et al. 1987; Longuet-Higgins 1981; Mitiche et al. 1985; Tsai and Huang 1984). This technique is reported to be noise sensitive even when more than two frames are used (Bharwani et al. 1986; Ullman 1983). Nonrigid motion, or the presence of several moving objects in the field of view, would cause a relatively large residual error in the solution of the system of equations. Moreover, in some cases of nonrigid motion, an acceptable numeric solution may exist that corresponds to a rigid motion interpretation. In such situations, the movements of individual entities in the field of view would not be detectable by the classic scheme. Adiv (1985) generalized this approach to handle multiple moving objects by using a complex grouping process to segment the optical flow field.

For applications with mainly translational camera movements, such as robotic land vehicles, alternative approaches have been developed to make use of this particular form of self-motion (Jerian and Jain 1984; Longuet-Higgins and Prazdny 1980; Prazdny 1981). To reconstruct the 3D scene structure, some researchers have assumed planar motion (Marimont 1986) or even pure camera translation (Bolles and Baker 1985; Jain 1983; Lawton 1983). Usually, unlike our scenario, a completely static environment is assumed. An important concept 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 camera translation in the forward direction). In practice, locating the FOE accurately is generally impossible under noisy conditions. We have therefore extended this concept by computing a patch of possible FOE locations, called the *Fuzzy FOE*, instead of a single point.

The emphasis of this paper is on the application of qualitative techniques for motion understanding, with the Fuzzy FOE and the accompanying reasoning process as its main components.

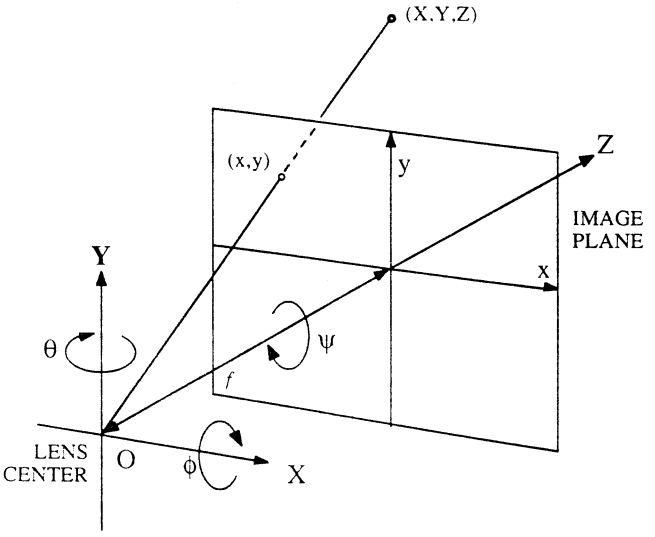
Our approach has two main components. Given a set of point correspondences (Barnard and Thompson 1980; Moravec 1977) for each pair of frames, we first compute the Fuzzy FOE and remove the effects of camera rotation. In the second step, we use the 2D locations and motion of features relative to each other and relative to the Fuzzy FOE to reason about the 3D scene structure as well as 3D motion. These results are used to incrementally construct a model of the environment that includes the information about the static scene structure and the moving objects therein. This reasoning process and the scene model are characterized by two key features: the emphasis on qualitative techniques and the ability to pursue multiple interpretations simultaneously.

It is to be noted that difficulties in separating rotation and translation components may arise when the camera is directed perpendicular (or almost perpendicular) to the direction of vehicle heading. For example, this is the case for a flat scene observed from a side-looking camera. In the described important practical application of robotics for land vehicles, where the camera looks approximately in the direction of vehicle motion, there problems will not occur. Our approach for computing the motion parameters relies on an error function that is, at least locally, well behaved. Several error functions were evaluated (Bhanu and Burger 1988), among them those suggested by other researchers. It turned out that the widely proposed technique of extending displacement vectors onto straight lines and evaluating their intersections leads to error functions that are not well behaved at all. The function we employ assumes that the FOE is given (i.e., hypothesized) and supplies the corresponding optimal rotations for that FOE in closed form. Consequently, our algorithm iterates over FOE locations (not rotations), using a local search strategy.

Although quantitative techniques have traditionally been dominant in machine vision, qualitative techniques are now receiving increasing attention in this field (Burger and Bhanu 1987; 1988; 1989; Thompson and Kearney 1986; Verri and Poggio 1987). They hold the potential to replace expensive numeric computations and models (with often unnecessary precision) by a simpler process that reasons about the important properties of the scene, using less precise representations. This is particularly true for the higher levels of vision but seems to be a useful path for building abstract descriptions gradually, starting at the lowest level of vision. Multiple scene interpretations are supported to reflect the ambiguities inherent to any type of scene analysis. If only one interpretation was 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, an appropriate interpretation (e.g., the most "plausible" or the most "threatening" interpretation) can be selected.

The overall process of constructing the scene interpretations consists of three main steps. First, significant features (points, boundaries, corners, etc.) are extracted from the image sequence, and the 2D displacement vectors are computed for each frame pair. In the following, we employ only point features and assume that the problem of selecting and matching corresponding points is solved (Barnard and Thompson 1980; Kim and Bhanu 1987). In the second step, we use the original displacement field to compute the Fuzzy FOE (i.e., the vehicle's approximate direction of heading and the amount of rotation in space). Most of the necessary quantitative computations are performed in this 2D step, which is described in section 2. The third step (2D Change Analysis) constructs the 3D *Qualitative Scene Model* by analyzing the movements of individual features with respect to the Fuzzy FOE location (section 3). Experiments with our approach on real images taken from the Autonomous Land Vehicle (ALV) are discussed in section 4, and section 5 presents the conclusions of the paper.

Fig. 1. Camera-centered coordinate system. The origin of the coordinate system is located at the lens center of the camera. The focal length f *is the distance between the lens center and the image plane. A 3D point (X Y Z) is mapped onto the image location (x y).*



2. The Fuzzy FOE

When a camera undergoes pure forward translation along a straight line in space, the images of all stationary features seem to diverge out of one particular location in the image, commonly called the "focus of expansion" (FOE). In reality the vehicle not only translates but also rotates about its three major axes. For our purpose, the movement M of a land vehicle can be sufficiently approximated by a translation $\mathbf{T}$ followed by rotations about the vertical axis $\mathbf{R}_\theta$ (pan) and the horizontal axis $\mathbf{R}_\phi$ (tilt), ignoring the yaw component $\mathbf{R}_\psi$. A 3D point $\mathbf{X} = (x, y, z)$ in the camera-centered coordinate frame (Fig. 1) is thus transferred by the camera movement M to a new location $\mathbf{X}' = (x', y', z')$

$$M: \mathbf{X} \rightarrow \mathbf{X}' = \mathbf{R}_\phi \, \mathbf{R}_\theta \, \mathrm{T} \, (\mathbf{X}).$$

If the observed scene is completely stationary, the effects on the image caused by the camera movement M can be described by a 2D transformation d (for *displacement*), which takes the original image I to the subsequent image I'. The 3D rotations $\mathbf{R}_\varphi$ and $\mathbf{R}_\theta$ and translation $\mathbf{T}$ have their equivalents in d as the separate 2D transformations $\mathbf{r}_\varphi$, $\mathbf{r}_\theta$, and $\mathbf{t}$:

$$d: I \rightarrow I' = \mathbf{r}_\varphi \, \mathbf{r}_\theta \, \mathbf{t} \, (I).$$

Fig. 2. Interpretation of the Focus of Expansion (FOE). Vehicle motion between its initial position (where image I is observed) and its final position (image I') is modeled as two separate steps. First the vehicle translates by a 3D vector T from position P_0 to position P_1 without changing its orientation Ω_0 **(A).** *After this step, the intermediate image I* would be seen. Subsequently* **(B),** *the vehicle rotates by changing its orientation form Ω_0 to Ω_1. Now image I' is observed. The FOE is found where the vector T intersects the image plane I (and also I*).*

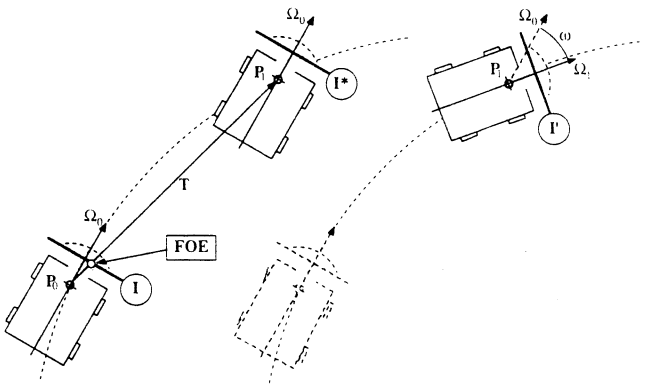
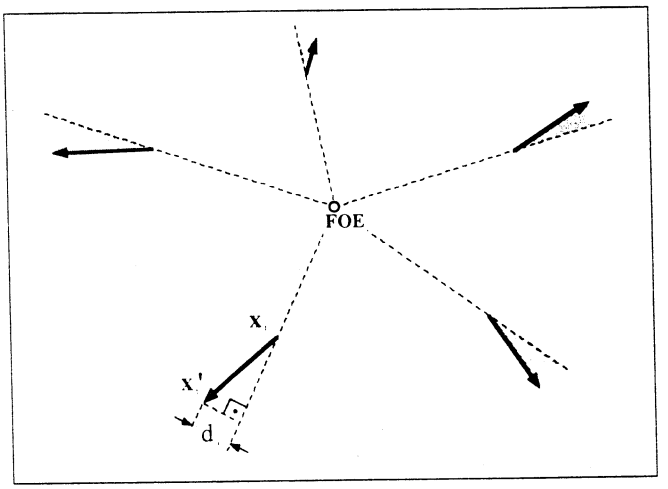


Fig. 3. Measuring the deviation from a radial expansion pattern. For a hypothetical FOE and a given set of displacement vectors $\{x_i \rightarrow x'_i\}$, the deviation is defined as the sum of the perpendicular distances Σd_i.



Ignoring the effects at the boundary of the image, as pure camera rotations do not supply new aspects of the 3D environment, the corresponding 2D transformations $\mathbf{r}_\varphi$ and $\mathbf{r}_\theta$ are effectively the mappings of the image onto itself. Conversely, the image effects **t** of pure camera translation depend on each 3D point's actual location in space. We introduce an (hypothetical) intermediate image *I**, which is the result of a pure camera translation **T**:

$$\mathbf{t}: I \rightarrow I^*.$$

Notice that the image *I** is never really observed, except in the special case of pure camera translation (Fig. 2). However, *I** has two important properties. First, all displacement vectors between corresponding points in *I* and *I** seem to diverge from a particular image location (x_f, y_f) known as the FOE, unless the camera does not translate at all. We call this property of the displacement field "radial_mapping (I, I^*)." *Second*, for given tilt and pan angles φ and θ, *I** can be obtained regardless of the 3-D scene structure by applying the inverse mappings $\mathbf{r}_\varphi^{-1}$ and $\mathbf{r}_\theta^{-1}$ (which always exist) to the observed image *I'*:

$$I^* = \mathbf{r}_\theta^{-1}\, \mathbf{r}_\varphi^{-1}\, I'.$$

Once suitable mappings $\mathbf{r}_\theta^{-1}\, \mathbf{r}_\varphi^{-1}$ have been found, the FOE can be located for the pair of images *I* and *I**. However, it is not trivial to determine how close a given displacement field is to a radial mapping without knowing the location of the FOE. In most of the proposed schemes for testing this property, the displacement vectors are extended as straight lines to measure the spread of their intersections (Jerian and Jain 1984; Prazdny 1983). Unfortunately, the resulting error functions are noise sensitive and not well behaved for varying values of φ and θ (i.e., they require expensive global search) (Bhanu and Burger 1988; Burger and Bhanu 1989).

Alternatively, we can hypothesize a particular FOE and then measure how the displacement field resembles a radial pattern emanating from this FOE. The sum of the perpendicular distances between radial rays and the end points of the displacement vectors is a simple and useful measure (Fig. 3). The optimal rotation angles for a particular FOE (i.e., those that would minimize this deviation) and the remaining error can be found analytically. This remaining error is used as the criterion to evaluate a hypothetical FOE. When plotted as a 2D distribution, the resulting error function is smooth and monotonic within a large area around the actual FOE. This means that even from a poor initial guess the global optimum can be found by local search methods, such as steepest descent. A detailed derivation of this error function and its behavior under noisy conditions can be found in Bhanu and Burger (1988) and Burger and Bhanu (1989).

Although this technique is robust even in the pres-

ence of considerable noise and under small camera translation, the 2D error function flattens out in these extreme cases and the location of minimum error may be considerably off the actual FOE. The local shape of the error function is therefore an important indicator for the accuracy of the result. This raises the question of whether it is reasonable to locate the FOE as one particular point in the image. After all, even humans seem to have difficulties in estimating the direction of heading under similar conditions (Rieger and Lawton 1985).

We have therefore extended the concept of the FOE to specify not a single image location but a connected region, termed the Fuzzy FOE, that reflects the shape of the error distribution. In general, a flat error function is reflected by a large Fuzzy FOE (i.e., little accuracy in the location of the FOE), whereas a small region indicates a distinct local optimum for the FOE. The following algorithm computes the Fuzzy FOE by first looking for the bottom of the error function and then accumulating surrounding FOE-locations (see Fig. 3).

Fuzzy-FOE (I, I'):

(Compute the Fuzzy FOE for a given pair of images I and I').

(1) Guess initial FOE (x_0, y_0) (e.g., the FOE obtained from the previous frame pair) and compute the corresponding optimal rotations φ_0, θ_0 and the deviation from a radial flow field error e_0.
(2) From (x_0, y_0) start a local search (e.g., steepest descent) for an FOE location (x_c, y_c) that results in a mimimun error e_c.
(3) Create the set FUZZY-FOE = $\{(x_c, y_c, e_c)\}$.
(4) Grow the set FUZZY-FOE by including adjacent FOE-locations (x_i, y_i, e_i) until the accumulated error $E = e_c + \Sigma e_i$ exceeds a predefined limit.

After computing the Fuzzy FOE and the angles of horizontal and vertical rotation, a good estimate for the motion parameters of the vehicle is available. Notice that this is possible without knowing the 3D structure of the observed scene. Also it is to be noted that to measure the camera motion with respect to the stationary world, none of the displacement vectors used for this computation may belong to another moving object. This information is supplied by the internal scene model (as described in the following section), which, among other things, tells us what features are currently believed to be stationary.

Fig. 4. Fuzzy FOE for a simulated displacement field. The vehicle is translating forward and rotating to the right by 2°. The small square in the center is the location of the actual FOE. The error values for surrounding (i.e., hypothesized) FOE locations are shown with circles of proportional size. Notice the elongated shape of the FOE region, which is a result of the particular distribution of displacement vectors (typical for road scenes).

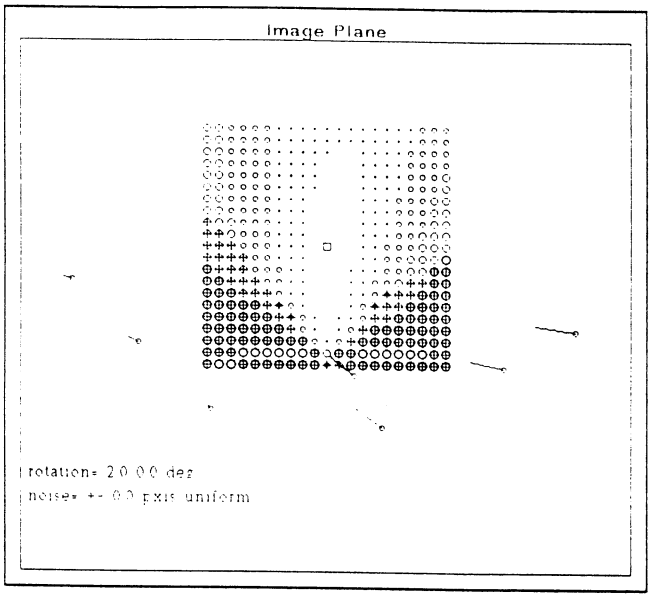


Fig. 4 shows the results of applying this algorithm to a simulated sparse displacement field. The shape of the error function around the actual FOE is plotted with circles of size proportional to the error. The blank area in the center of Fig. 4 marks the resulting Fuzzy FOE. A detailed error analysis for this particular FOE algorithm can be found in Bhanu and Burger (1988).

Our approach is primarily designed to be used for a large number of important practical applications requiring a forward-looking camera. Separating the motion components becomes increasingly difficult when the FOE moves far off the optical axis. The FOE, however, is not required to lie within the bounds of the image.

3. Constructing a Qualitative Scene Model

The choice of a suitable scheme for the internal representation of the scene is of great importance. It is as-

sumed that all feature points used for the FOE computation belong to the stationary environment. This assumption is part of the *Qualitative Scene Model* (*QSM*). The QSM is a 3D camera-centered interpretation of the scene that is built incrementally from visual information gathered over time. The nature of this model, however, is *qualitative* rather than a precise geometric description of the scene. The basic building blocks of the QSM are *entities*, which are the 3D counterparts of the 2D *features* observed in the image. For example, the point feature A located in the image at x, y at time t, denoted by (POINT-FEATURE $At \times y$), has its 3D counterpart in the model as (POINT-ENTITY A).

Because the model is camera centered, the image locations and 2D movements of features are implicitly part (i.e., known facts) of the model. Additional entries are the properties of entities (e.g., "stationary" or "mobile") and relationships between entities (e.g., "closer"), which are not given facts but are the outcome of some interpretation step (i.e., hypotheses). The hypotheses are expressed in the model as either

(STATIONARY *entity*) or (MOBILE *entity*).

The assertion (STATIONARY x) is really a hypothesis about entity x in the 3D environment that may be either true or false. By default, newly observed entities are classified as stationary. This hypothesis is withdrawn as soon as there is any indication (from image observations) that this entity could be actually mobile.

A key feature of the QSM is that it generally contains not only one interpretation of the scene but a (possibly empty) *set* of interpretations that are all pursued simultaneously. At any point in time, a hypothesis is said to be "feasible" if it exists in the QSM and does not conflict with some observation made since it was established.

Interpretations are structured as an inheritance network of partial hypotheses. Individual scene interpretations are treated as "closed worlds," i.e., a new conclusion only holds within an interpretation where all the required premises are true. Interpretations are also checked for internal consistency (e.g., entities cannot be both stationary and mobile within the same interpretation). The QSM is maintained through a generate-and-test process as the core of a rule-based blackboard system. The two major groups of rules are: Generation Rules and Verification Rules (Fig. 5). In the following discussion, we use the original notation of the *ART* (Clayton 1985) language for defining rules. Fig. 5 shows the overall structure of the interpretation process. A large portion of the rules are derived directly from the laws of perspective imaging. The rules that reflect some form of heuristics (which hold for a large class of scenes in practical applications) are clearly marked.

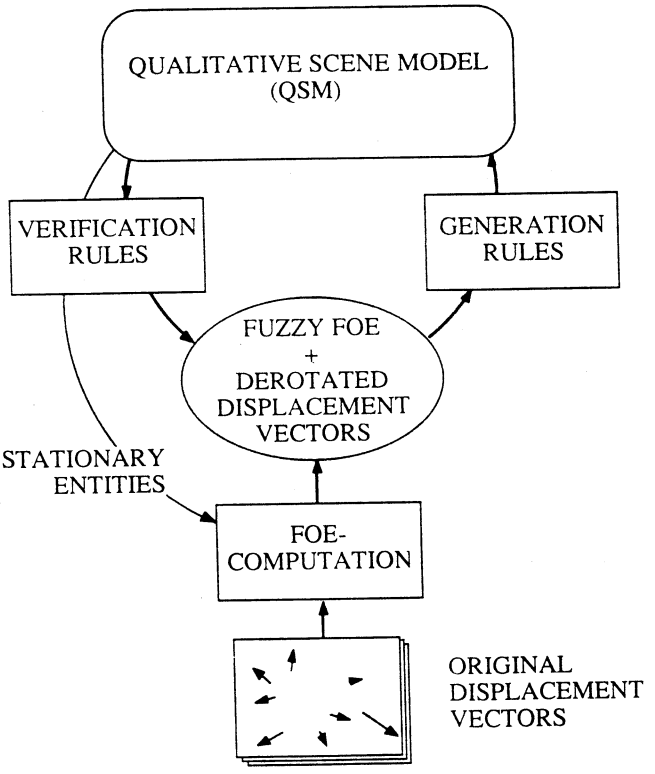


Fig. 5. Overall structure of the interpretation process. From the original displacement vectors (obtained by matching corresponding features), the Fuzzy FOE and the derotated displacement field are computed. The Qualitative Scene Model (QSM) is built in a hypothesize-and-test cycle by two sets of rules. Generation rules search for significant image events and place immediate conclusions (hypotheses) in the model. Verification rules check existing hypotheses if they are consistent with the changes occurring in the image. A set of environmental entities that are believed to be stationary is supplied by the QSM for their use in the FOE computation.

3.1. Generation Rules

Generation rules examine the (derotated) image sequence for significant changes and modify each inter-

pretation in the QSM if applicable. Some of these observations have unconditional effects on the model (e.g., if an image feature is found to be moving *toward* the Fuzzy FOE, instead of moving away from it, then it belongs to a moving entity in 3D space). The actual rule contains only one premise and asserts (MOBILE ?x) as a globally known fact (i.e., one that is true in every interpretation):

```
(defrule DEFINITE-MOTION
  (MOVING-TOWARD-FOE ?x ?t)        <observation at time t>
  ⇒
  (assert (MOBILE ?x))).           <a global fact>
```

Similarly, if two image features A and B lie on opposite sides of the Fuzzy FOE and they are getting closer to each other, then they must be in relative motion in 3-D space:

```
(defrule RELATIVE-MOTION
  (OPPOSITE-FOE ?x ?y ?t)          <image observation 1 (global)>
  (CONVERGING ?x ?y ?t)            <image observation 2 (global)>
  ⇒
  (assert (MOVEMENT-BETWEEN ?x ?y))).   <a new global fact>
```

Other observations depend on the facts that are currently true within a "world" and therefore may have only local consequences inside particular interpretations. The following rule pair responds to the new fact created by the above rule by creating two new hypotheses. If an interpretation exists that considers at least one of the two entities (x, y) stationary, then the other entity cannot be stationary (i.e., it must be mobile):

```
(defrule RELATIVE-MOTION-X
  (MOVEMENT-BETWEEN ?x ?y)         <a global fact>
  (STATIONARY ?x)                  <true only inside an interpretation>
  ⇒
  (assert (MOBILE ?y)))            <new fact local to this interpretation>
```

```
(defrule RELATIVE-MOTION-Y
  (MOVEMENT-BETWEEN ?x ?y)         <a global fact>
  (STATIONARY ?y)                  <true only inside an interpretation>
  ⇒
  (assert (MOBILE ?x)))            <new fact local to this interpretation>
```

Although some image observations allow direct conclusions about motion in the scene, other observations give clues about the stationary 3D structure. If the *exact* location of the FOE is known, then the depth of each stationary point (i.e., its 3D distance from the camera) is proportional to the rate of divergence (from the FOE) of that point (Prazdny 1983). Applied to the Fuzzy FOE, where a set of potential FOE locations is given, the distance $Z(A)$ of a stationary point A is determined as an *interval* instead of a single number:

$$Z^{\min}(A) \leq Z(A) \leq Z^{\max}(A).$$

Therefore point A must be closer in 3D than another point B if the corresponding ranges of depth do not overlap, i.e.,

$$Z^{\max}(A) < Z^{\min}(B) \Rightarrow (\text{CLOSER } A\ B).$$

Since this conclusion only holds if both entities are actually stationary, the following rule fires only within a suitable interpretation (if it exists):

```
(defrule CLOSER-FROM-DIVERGENCE
  (STATIONARY ?x)                  <interpretation where both x and y are stationary>
  (STATIONARY ?y)
  (test (< (Zmax ?x) (Zmin ?y)))   <no overlap in range>
  ⇒
  (assert (CLOSER ?x ?y))).        <a new hypothesis>
```

To compare the ranges of 3D points, another criterion can be used that does not measure the individual rate of divergence. According to this criterion, the change of distances *between* certain pairs of features is observed. If two stationary points lie on the same side of the FOE and the distance between them is becoming smaller, then the *inner* feature (i.e., the one that is

nearer to the FOE) is closer in 3D space. This test is valuable for features that are relatively close to each other. It can be employed even if the image is not (or incorrectly) derotated and the location of the FOE is either only known very roughly or is completely outside the field of view (i.e., for a side-looking camera):

```
(defrule CLOSER-FROM-CHANGING-DISPARITY
  (STATIONARY ?x)              <interpretation where both x
                                 and y are stationary>
  (STATIONARY ?y)
  (SAME-SIDE-OF-FOE ?x ?y)     <e.g. both are right of the FOE>
  (CONVERGING ?x ?y)           <dist. between x and y is de-
                                 creasing>
  (INSIDE ?x ?y)               <x is nearer to the Fuzzy FOE
                                 than y>
  ⇒
  (assert (CLOSER ?x ?y))).    <a new hypothesis>
```

Although the purpose of the generation rules is to establish new hypotheses and conclusions, the purpose of verification rules is to review interpretations after they have been created (Fig. 5) and, if possible, prove that they are false. When a hypothesis is found to be inconsistent with some new observation, that hypothesis is usually removed from the QSM. Simultaneously, any interpretation that is based on that hypothesis is removed. Because we are always trying to come up with a single (and hopefully correct) scene interpretation, this mechanism is important for pruning the search tree. Notice that all the rules described so far are based upon the known effects of perspective imaging (i.e., they are valid for any type of scene).

3.2. Verification Rules

Verification rules fall into two categories. One group of rules verifies the *internal* consistency of the scene model. For example, a particular entity cannot be labeled both stationary *and* mobile in one single interpretation. The following rule detects those cases and removes ("poisons") the affected hypothesis:

```
(defrule REMOVE-STATIONARY-AND-MOBILE
  (STATIONARY ?x)     <this is an inconsistent hypothesis>
  (MOBILE ?x)
  ⇒
  (poison)).          <remove this hypothesis>
```

Similarly, the CLOSER-relation may not be symmetric for any pair of stationary entities. For a non-symmetric situation, we conclude that there is some 3D movement between the two entities:

```
(defrule CHECK-FOR-CLOSER-SYMMETRY
  (CLOSER ?x ?y)      <this is an inconsistent hypothesis>
  (CLOSER ?y ?x)
  ⇒
  (at ROOT (assert (MOVEMENT-BETWEEN ?x ?y)))).    <a
new global fact>
```

The second group of verification rules checks whether existing hypotheses (created in the past) are compatible with the current activities in the image. Usually these rules, if used as generators, would produce a large number of unnecessary conclusions. For example, the general layout of the scene (observed from the top of a land vehicle) suggests the rule of thumb that things that are *lower* in the image are generally *closer* to the camera. Otherwise, some motion has probably occurred between the two entities involved. The first of the following rules signals that conflict and the other pair of rules creates two different hypotheses about the direction of motion:

```
(defrule LOWER-IS-CLOSER-HEURISTIC
  (CLOSER ?x ?y)      <existing hypothesis>
  (BELOW ?y ?x ?t)    <image observation: actually x should be
                        below y>
  ⇒
  (at ROOT (assert (LOW-CLOSE-CONFLICT ?x ?y ?t))))

(defrule CONCLUDE-RECEDING-MOTION
  (LOW-CLOSE-CONFLICT ?x ?y ?t)
  (STATIONARY ?x)
  ⇒
  (assert (MOBILE ?y) (MOVES-RECEDING ?y ?t)))
```

```
(defrule CONCLUDE-APPROACHING-MOTION
  (LOW-CLOSE-CONFLICT ?x ?y ?t)
  (STATIONARY ?y)
  ⇒
  (assert (MOBILE ?x) (MOVES-APPROACHING ?x ?t))).
```

In summary, the construction of the QSM and the search for the most plausible scene interpretation are 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.

The information contained in the QSM is useful for a variety of purposes. First it supplies a partial ordering in depth for the static entities in the scene, which is important in scene assessment and navigation. Threat analysis can be based on the mobile entities in the QSM. Finally, the FOE computation must be supplied with a set of features that are currently believed to be stationary (i.e., those that are not considered mobile in any existing scene interpretation).

Although perspective imaging has been the motivation for the rules described here, other important visual clues are available from occlusion analysis, perceptual grouping, and semantic interpretation. *Occlusion* becomes an interesting phenomenon when features of higher dimensionality than points are employed, such as lines and regions. Similarities in form and motion found by *perceptual grouping* allow us to assemble simple features into more complex aggregates. Finally, as an outcome of the recognition process, *semantic* information may help to disambiguate the scene interpretation. If an object has been recognized as a building, for example, it makes every interpretation obsolete that considers this object mobile. These are the central topics for future extensions.

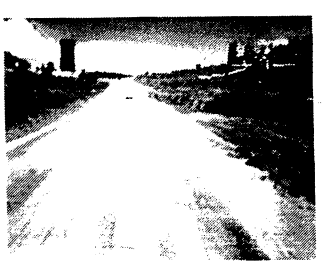
4. Experiments

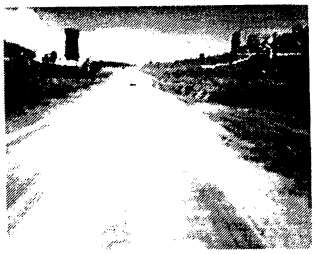
We have implemented a prototype system using the ART (Clayton 1985) expert system tool on a Symbolics 3670 computer. The FOE component was programmed with Common LISP functions. Low-level processing (edge detection) was done on a VAX 11/750. In the following the operation of the QSM and the associated rule base is demonstrated on an image sequence shown in Fig. 6. The image sequence was obtained from the Autonomous Land Vehicle (ALV) driving on a road at a test site in Colorado. The images contain two moving objects: a car that has passed the ALV and barely visible in the distance, and a second car that is approaching in the opposite direction and is about to pass. From the original sequence provided on a video tape with a frame rate of 30/s, images were taken in 0.5-s intervals (i.e., at a frame rate of 2/s). The images were digitized to a spatial resolution of 512×512, using only the Y-component (luminance) of the original color signal.

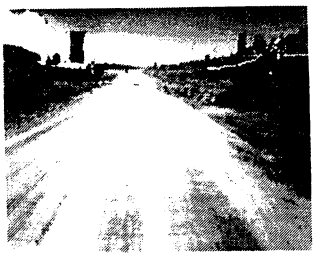
To obtain the displacement vectors, point features were selected and tracked manually between successive frames. Binary edge images were used to imitate the conditions for automatic point tracking, because some clues visible in the original grey-scale sequence are lost during edge detection. Consequently, the end points of the original displacement vectors are not very accurate. Recent experiments on extended sequences (Bhanu 1988; Kim and Bhanu 1987) show that similar results can be achieved with fully automatic feature tracking.

Figs. 7–9 show the edge images of 16 frames with the points being tracked labeled with ascending numbers. The actual image location of each point is the lower left corner of the corresponding mark. Points are given a unique label when they are encountered for the first time. After the tracking of a point has started, its label remains unchanged until this point is no longer tracked. When no correspondence is found in the subsequent frame for a point being tracked, be-

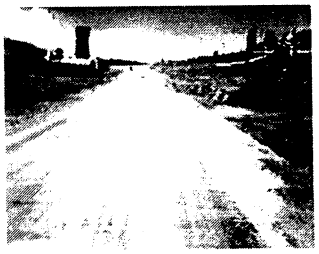
Fig. 6. Data used to evaluate qualitative reasoning and modeling (frames 182–197).


Frame 182

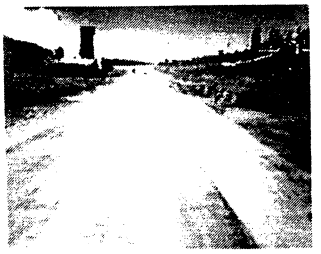

Frame 183

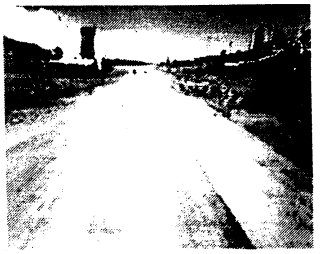

Frame 184


Frame 185

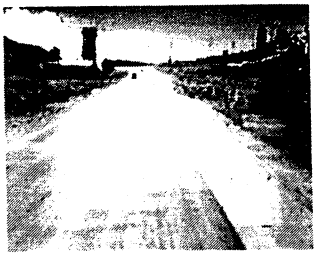

Frame 186

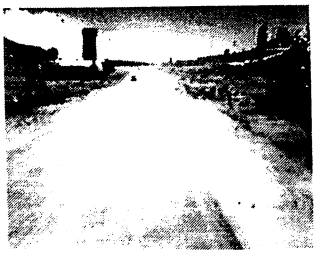

Frame 187

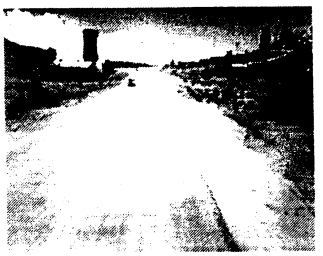

Frame 188


Frame 189

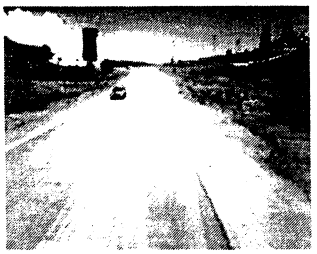

Frame 190

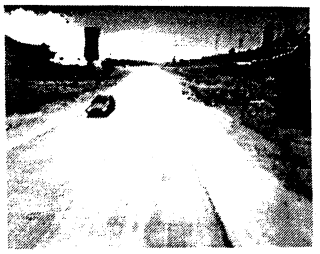

Frame 191


Frame 192


Frame 193


Frame 194

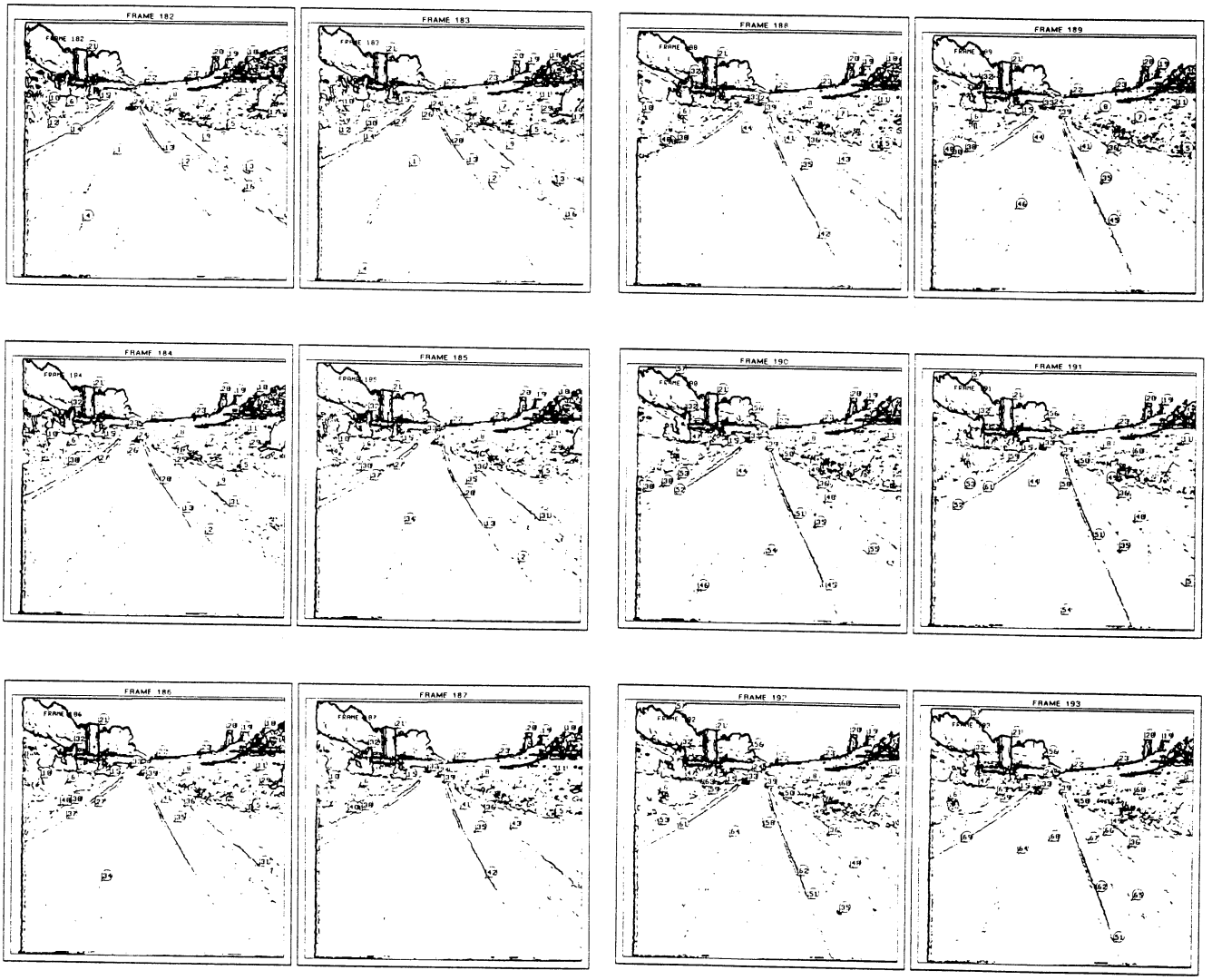

Frame 195


Frame 196


Frame 197

Fig. 7. Frames 182–187 of the original image sequence taken from the moving ALV after edge detection and point tracing. The points are located at the lower left corners of their marks. The scene contains two moving objects, one car moving away from the ALV (point 24) and another car approaching the ALV (point 33).

Fig. 8. Frames 188–193 of the original image sequence after edge detection and point tracing.



cause of occlusion, because the feature left the field of view, or because it could not be identified, tracking of this point is discontinued. Should the same point appear again, it is treated as a new item and given a new label. Approximately 25 points per image have been selected in the sequence shown in Fig. 6.

Figs. 10–12 show the original set of displacement vectors (solid lines) between pair of frames, the "Fuzzy" FOE (shaded area), and the "derotated" displacement vectors (dotted lines). The rotation scale in the lower left corner indicates rotation angles between the pair of frames. Only the stationary points are used to compute the FOE, vehicle rotation, and velocity.

The QSM processes the images and determines the motion of moving objects and builds a 3D representation of the environment as described in the last section. Figs. 13 through 18 show the complete scene interpretations starting at frame 182 up to frame 197. Interpretations are ranked by their number of stationary entities (i.e., "Interpretation 1" is ranked higher than "Interpretation 2" if both exist). During this run, the maximum number of concurrent interpretations was two. Whenever two interpretations exist at the

Fig. 9. Frames 194–197 of the original image sequence after edge detection and point tracing.

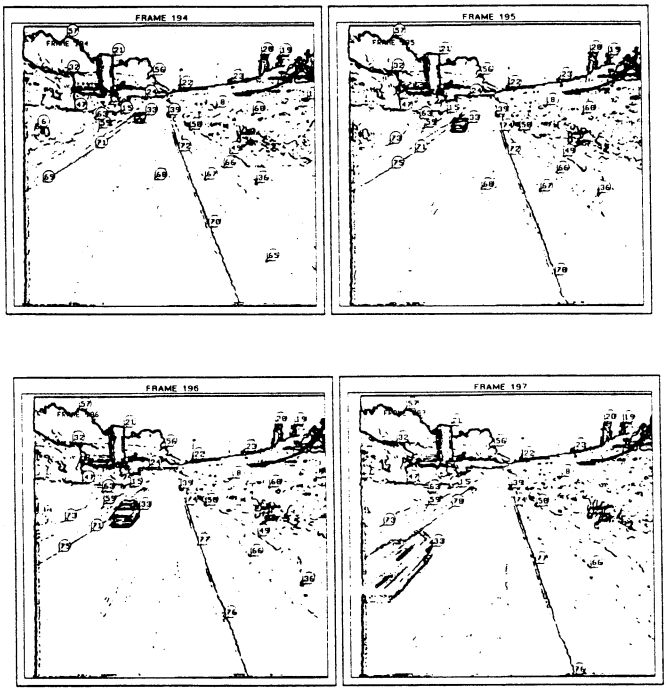
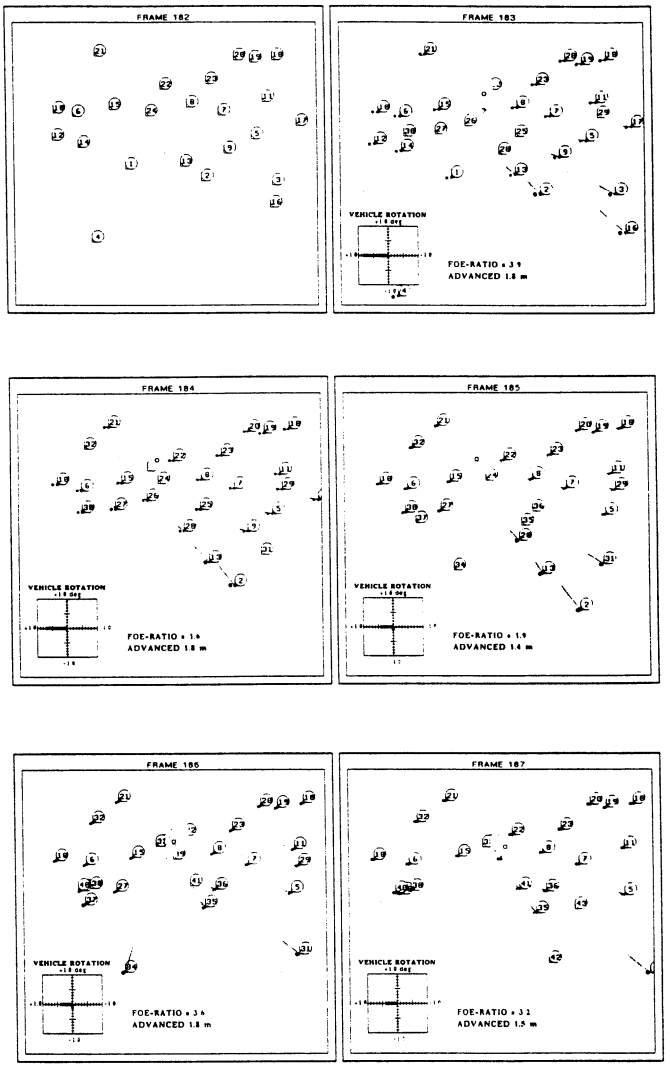


Fig. 10. Displacement vectors and estimates of vehicle motion for the image sequence shown in Fig. 7. Solid and dotted lines show original and derotated displacement vectors, respectively. Shaded area shows the Fuzzy FOE. The absolute advancement of the vehicle is estimated in meters. The vehicle rotation is plotted in a coordinate grid over ±1°.



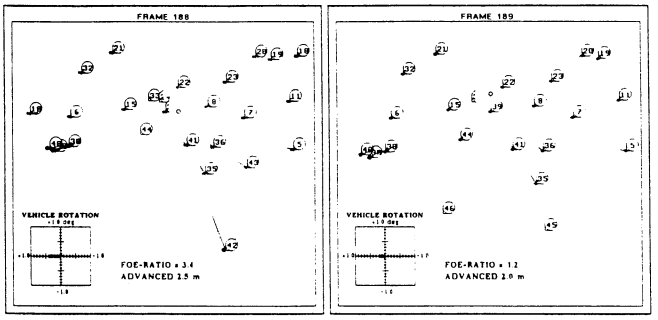
same time, they are lined up horizontally in the figures. Otherwise, interpretations are displaced to indicate that they refer to different points in time. Entities are marked as stationary or mobile. Entities that are considered stationary are marked with circles or plain labels. Arcs from a small circle (or plain label) to a larger circle indicate that a CLOSER-relationship has been established between the two entities. In these cases, the entity with the larger circle is considered closer to the camera in the 3D scene. Mobile entities are marked with squares if they are thought to be in some motion, or with arrows if the direction of their current movement has been determined.

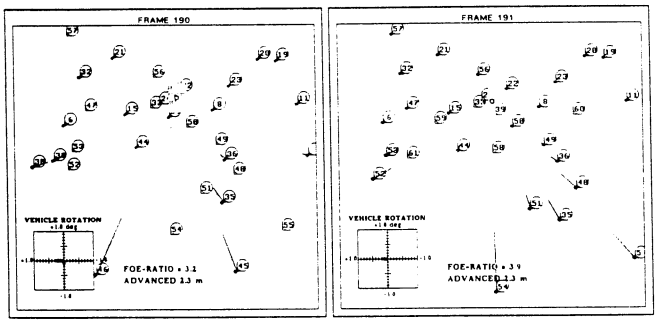
The scene contains two moving objects, a car (24) that has passed the ALV and is moving away throughout the sequence and another vehicle (33), approaching the ALV on the same road, that appears in frame 195 (Fig. 16).

Fig. 13 examines the state of the QSM at frames 182–184. The scene contains a number of stationary points and one moving point (24) that belongs to another vehicle that has passed the ALV and is moving away from the camera. First, the parameters of the ALV's self-motion are computed with respect to a set of environmental features believed to be stationary (Fig. 10). This set is defined by the hypotheses currently contained in the scene model and was described in section 3.

Fig. 13 visualizes two separate but feasible scene interpretations for the situation in frame pair 182–183. The existence of two interpretations is a result of the movement of the receding car (point 24). This movement was detected as 2D motion "across the FOE" (see rule RELATIVE_MOTION in section 3) between point 24 on one side of the FOE and points 8, 11, 19, 20, 22, 23 on the opposite side. Interpreta-

Fig. 11. Displacement vectors and estimates of vehicle motion for the image sequence shown in Fig. 8.





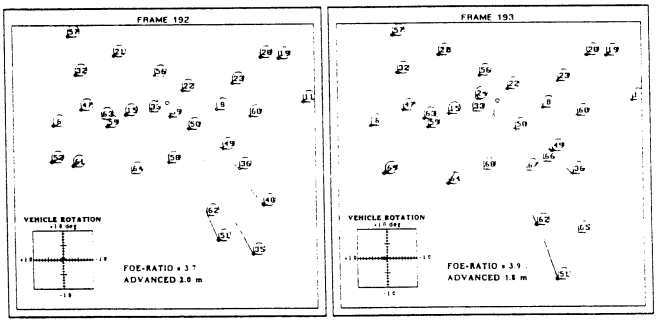
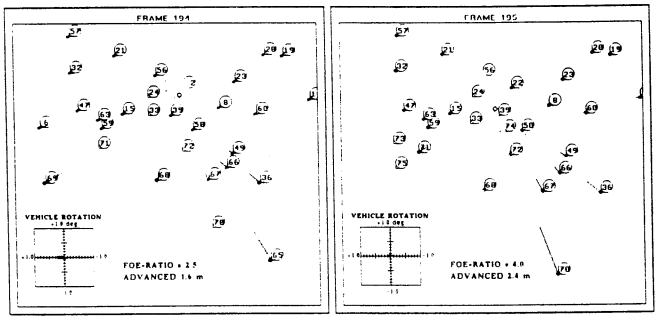
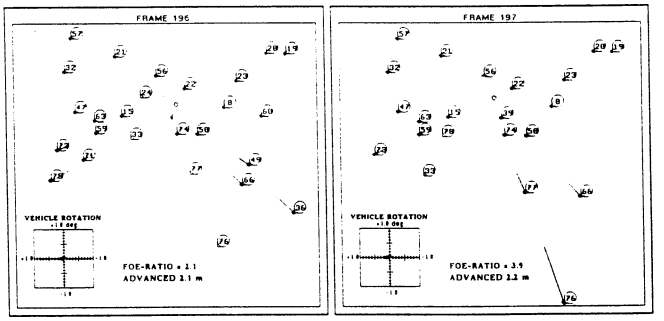


Fig. 12. Displacement vectors and estimates of vehicle motion for the image sequence shown in Fig. 9.





tion 1 considers all entities stationary, except point 24, which is moving upward (in the 3D coordinate frame). This corresponds to the actual situation. However, Interpretation 2 is also feasible, taking 24 as stationary and points 8, 11, . . . , 23 as moving downward. Notice that CLOSER-relationships are only formed between stationary entities.

In the subsequent frame pair (183–184, Fig. 13), point 24 is observed to move toward the FOE, which is a definite indicator for 3D motion relative to the camera (rule DEFINITE-MOTION in section 3). Any interpretation considering entity 24 as stationary is removed from the model, such that Interpretation 2 does not survive. Consequently, only one (correct) interpretation exists after frame 184.

A single interpretation is pursued from frame 184 until frame 194 (Figs. 14 to 16). In this period, no object motion other than the one caused by point 24 is observed. However, the perceived 3D structure of the stationary part of the scene is continuously refined by adding new CLOSER-relationships between entities. Point 24 is always considered mobile, although the direction of the movement cannot be identified between every pair of frames. After frame 195, two interpretations again become feasible, this time caused by the movement of the approaching car (point 33). The (correct) alternative 1 was ranked higher because of the larger number of stationary entities.

For frame 196 (pair 195–196), two feasible scene interpretations are created (Fig. 17), caused by the behavior of feature 33. This point belongs to another vehicle that is approaching the ALV on the same road. The first vehicle (point 24) is declared as mobile in both interpretations, but the direction of movement is

Fig. 13. Scene interpretations for image sequence shown in Fig. 6. Receding object (frames 182–184). Two different scene interpretations are created, caused by relative image motion (across the FOE) between points 24 (the receding car) on one side and points 8, 11, . . . , 23 on the other side. Scene interpretation 1: entity 24 (arrow) is considered mobile with upward motion in the image; all others are stationary. Scene interpretation 2: entities 8, 11, . . . , 23 move downward, and 24 is stationary. In the following frame pair, point 24 is observed moving toward the FOE such that it is definitely in motion. Any interpretation with (STATIONARY 24) can be eliminated, and only one interpretation survives.

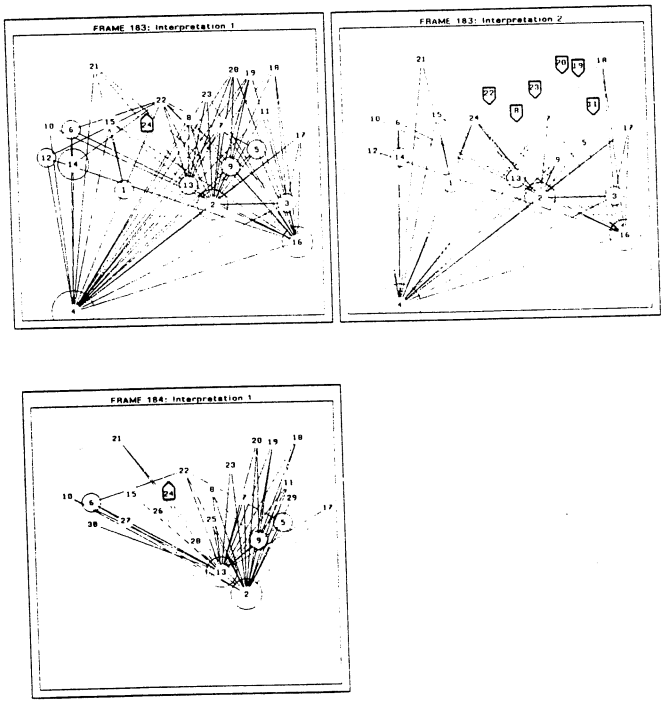
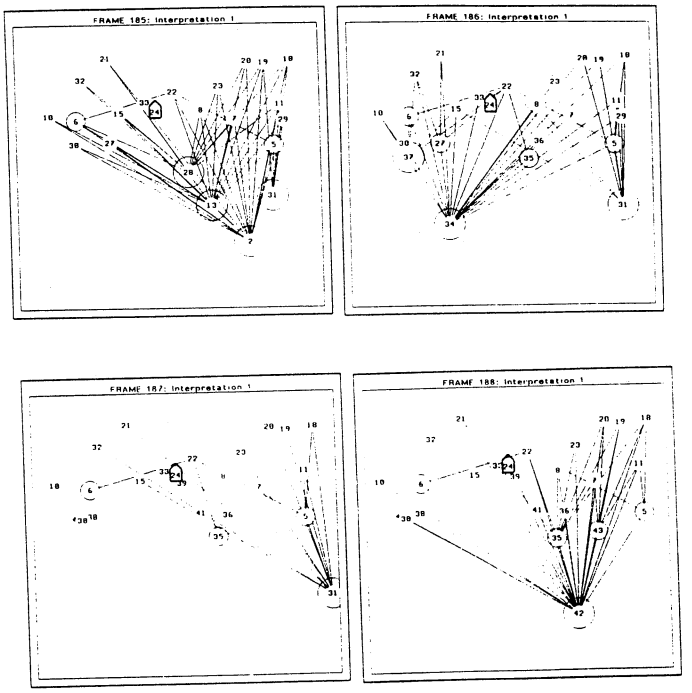


Fig. 14. Frames 185–188. The single interpretation from frame 184 is pursued because no object motion other than the one caused by point 24 is observed in this period. The 3D structure of the stationary part of the scene is continuously refined by adding new CLOSER-relationships between entities.



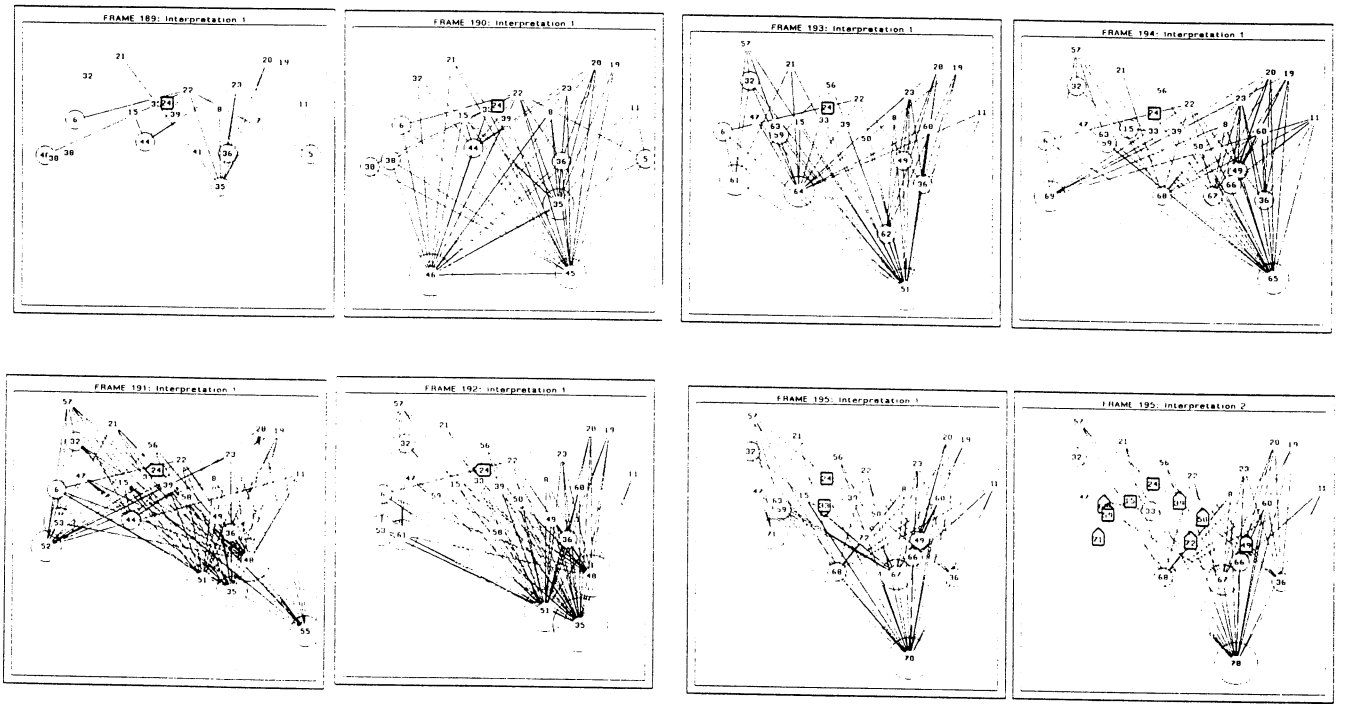
currently not known (indicated by a simple square). Some movement between feature 33 and several other features (15, 39, 50, . . . , 73) has been detected. This time, however, the direction of motion cannot be identified. Again two interpretations are created, with entity 33 labeled as mobile (Interpretation 1) or stationary (Interpretation 2). Both interpretations are carried over to frame 197, where two significant events happen.

In Interpretation 1 (Fig. 17), 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 the ALV." In Interpretation 2, entity 33 is still explained as stationary, as was the case in interpretation 2 of the previous frame. If this fact is 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 in section 3) and therefore Interpretation 2 can be ruled out. Only the correct interpretation (Fig. 18) remains.

The example illustrates the fact that some forms of object motion are difficult or even impossible to detect from single point features. For example, an entity approaching the camera along a straight line parallel to the viewing direction cannot be distinguished from a stationary point when both the camera and the object move uniformly. Such a situation occurred in frames 195–197, where another car approached the ALV. There we used heuristic reasoning about the general spatial layout of the scene to detect this motion indirectly. This experiment shows that the qualitative scene model (QSM) is robustly maintained under real-world conditions. The number of simultaneous interpretations is quite small (maximum is two), and the correct interpretation clearly ranks higher at any point in time.

Fig. 15. Frames 189–192. Point 24 is the only entity considered mobile in this period. Whenever the direction of motion could not be identified between a pair of frames, the mobile point was marked by a simple square. Notice the increased number of CLOSER-relationships established across the FOE area by different rates of expansion away from the FOE.

Fig. 16. Frames 193–195. After frame 195, two interpretations again become feasible, this time caused by the movement of the approaching car (point 33). The (correct) alternative 1 is ranked higher because of the larger number of stationary entities.



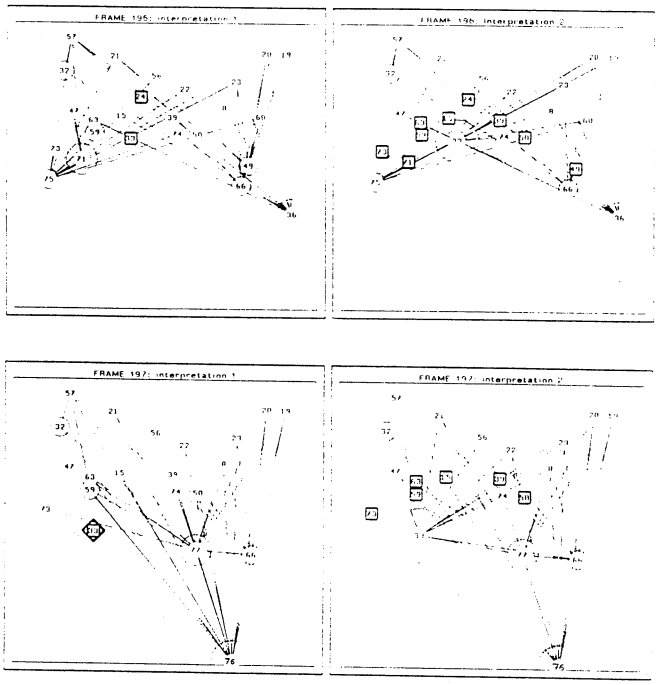
5. Conclusions

In this paper we have presented the conceptual outline of a new approach to scene understanding for mobile robots operating in dynamic environments. The challenge of understanding such image sequences is that stationary objects are generally not still in the image, and mobile objects do not necessarily appear to be in motion. Consequently, the detection of 3D motion sometimes requires reasoning far beyond simple 2D change analysis.

The approach taken here departs from previous related work by following a strategy of qualitative rather than quantitative reasoning and modeling. The numeric effort is packed into the computation of the Focus of Expansion (FOE), a low-level process that is performed entirely in 2D. We have extended the FOE concept to cope with the problems of noise and errors in the original displacement vectors. Instead of a single FOE, we determine a connected *region* of possible FOE locations, called the *Fuzzy FOE*, whose shape is directly related to the "goodness" of the displacement field.

The availability of reliable displacement vectors is vital to our approach. Although we used manual point tracking for the examples shown here, recent experiments indicate that automatic feature selection and tracking have become practical (Bhanu 1988; Kim and Bhanu 1987). All the interpretations and Fuzzy FOE results were generated by machine. Details on the FOE computation can be found elsewhere (Bhanu and Burger 1988; Burger and Bhanu 1989); the emphasis of this paper was to demonstrate how even a Fuzzy FOE can be used to draw powerful conclusions about motion and the 3D scene structure. A large part of the given rules follows directly from the laws of perspective imaging, although no rigorous proofs are given here. The rules that reflect some form of heuristics (which hold for a large class of scenes in practical applications) are clearly marked. From these clues, we construct and maintain an internal 3D representation, termed the *Qualitative Scene Model*, in a generate-and-test cycle over extended image sequences. To overcome the ambiguities inherent in dynamic scene

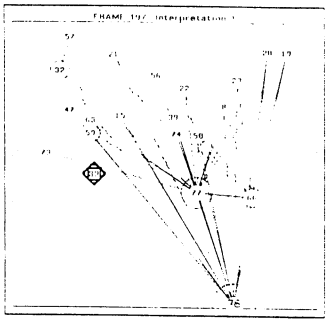
Fig. 17. Approaching object, frames 196–197. Two different scene interpretations are pursued simultaneously until frame 197. Entity 24 is known to be moving (from earlier conclusions) in both interpretations, but its direction of motion is currently undetermined (indicated by a square). Interpretation 1: entity 33 (square) is considered mobile with undetermined motion. Interpretation 2: entities 15, 39, 50, . . . , 73 (squares) are mobile, 33 is stationary. None of these interpretations can currently be ruled out and are carried over to the next frame pair. The receding car is not observed after frame 196. In the following frame pair, Interpretation 1: entity 33 is concluded to be moving towards the camera (indicated by an upright square). Interpretation 2 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 things lower in the image are generally closer in 3D space, which makes the entire interpretation 2 implausible.



analysis, multiple interpretations of the scene are pursued simultaneously. This model could also serve as a platform for other visual processes such as occlusion analysis, perceptual grouping, and object recognition.

For our implementation, we have used an off-the-shelf expert system tool mainly because it facilitates easy management of explicit knowledge, whereas speed was only of minor importance. The examples given in the text show the basic operation of our approach on real images acquired by the Autonomous Land Vehicle. The examples also demonstrate that some apparently simple situations require complex paths of reasoning. The system described here has been successfully applied to ALV sequences with over 250 frames in a fully automatic mode (Bhanu 1988).

Fig. 18. Final scene interpretation for frame 197. Point 33 has been identified correctly as approaching the camera. All other entities of the scene are considered stationary.



References

Adiv, G. 1985. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-7(4):384–401.

Barnard, S. T., and Thompson, W. B. 1980. Disparity analysis of images. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-2(4):333–340.

Bhanu, B. 1988. Knowledge based analysis of scene dynamics for target motion detection, recognition and tracking. DARPA project report (contract no. DACA 76-86-C-0017). Minneapolis, Minn.: Honeywell Systems and Research Center, pp. 2–68.

Bhanu, B., and Burger, W. 1988 (Cambridge, Mass., April). Qualitative motion detection and tracking of targets from a mobile platform. *Proc. DARPA Image Understanding Workshop,* pp. 289–318.

Bharwani, S., Riseman, E., and Hanson, A. 1986 (May). Refinement of environmental depth maps over multiple frames. *Proc. IEEE Workshop on Motion: Representation and Analysis,* pp. 73–80.

Bolles, R. C., and Baker, H. H. 1985 (October). Epipolar-plane analysis: A technique for analyzing motion sequences. *Proc. IEEE Workshop on Motion: Representation and Control,* pp. 168–178.

Bruss, A. R., and Horn, B. K. P. 1983. Passive navigation. *Comp. Vision Graph. Image Proc.* 12:3–20.

Burger, W., and Bhanu, B. 1987 (Milan, Italy, August).

Qualitative motion understanding. *Proc. Tenth International Joint Conference on Artificial Intelligence.* Los Altos, CA: Morgan Kaufmann Publishers, pp. 819–821.

Burger, W., and Bhanu, B. 1988 (Ann Arbor, Mich., June). Dynamic scene understanding for autonomous mobile robots. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 736–741.

Burger, W., and Bhanu, B. 1989 (San Diego, Calif., June). On computing a 'fuzzy' focus of expansion for autonomous navigation. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 563–568.

Clayton, B. D. 1985. *ART Programming Manual.* Los Angeles: Inference Corp.

Faugeras, O. D., Lustman, F., and Toscani, G. 1987 (London, June). Motion and structure from point and line matches. *Proc. First International Conference on Computer Vision,* pp. 25–34.

Jain, R. 1983. Direct computation of the focus of expansion. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-5(1):58–64.

Jerian, C., and Jain, R. 1984. Determining motion parameters for scenes with translation and rotation. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-6(4):523–530.

Kim, J., and Bhanu, B. 1987. Motion disparity analysis using adaptive windows. Technical report 87SRC38, contract no. DACA 76-86-C-0017. Minneapolis, Minn.: Honeywell Systems and Research Center.

Lawton, D. T. 1983. Processing translational motion sequences. *Comp. Vision Graph. Image Proc.* 22:114–116.

Longuet-Higgins, H. C., and Prazdny, K. 1980. The interpretation of a moving retinal image. *Proc. R. Soc. London* B 208:385–397.

Longuet-Higgins, H. C. 1981. A computer algorithm for reconstructing a scene from two projections. *Nature* 293:133–135.

Marimont, D. H. 1986 (May). Projective duality and the analysis of image sequences. *Proc. IEEE Workshop on Motion: Representation and Analysis,* pp 7–14.

Mitiche A., Seida S., and Aggarwal, J. K. 1985 (San Francisco, June). Determining position and displacement in space from images. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 504–509.

Moravec, H. P. 1977 (August). Towards automatic visual obstacle avoidance. *Proc. Fifth International Joint Conference on Artificial Intelligence,* p. 584.

Nagel, H.-H. 1986 (Paris, October). Image sequences—ten (octal) years—from phenomenology towards a theoretical foundation. *Proc. Int. Conference on Pattern Recognition,* pp. 1174–1185.

Prazdny, K. 1981. Determining the instantaneous direction of motion from optical flow generated by a curvilinear moving observer. *Comp. Vision Graph. Image Proc.* 17:238–259.

Prazdny, K. 1983. On the information in optical flows. *Comp. Vision Graph. Image Proc.* 22:239–259.

Rieger, J. H., and Lawton, D. T. 1985. Processing differential image motion. *J. Opt. Soc. Am.* A 2(2), pp. 354–360.

Thompson, W. B., and Kearney, J. K. 1986 (May). Inexact vision. *Proc. IEEE Workshop on Motion: Representation and Analysis,* pp. 15–21.

Tasi, R. Y., and Huang, T. S. 1984 (January). Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-6(1):13–27.

Ullman, S. 1983. Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion. Cambridge, Mass., A. I. Memo No. 721, MIT Artificial Intelligence Lab.

Verri, A., and Poggio, T. 1987 (Los Angeles, February). Qualitative information in the optical flow. *Proc. DARPA Image Understanding Workshop,* pp. 825–834.