# A Framework for Visual Motion Understanding

JOHN K. TSOTSOS, JOHN MYLOPOULOS, H. DOMINIC COVVEY, MEMBER, IEEE, AND STEVEN W. ZUCKER, MEMBER, IEEE

*Abstract*—A framework for the abstraction of motion concepts from sequences of images by computer is presented. The framework includes:

1) representation of knowledge for motion concepts that is based on semantic networks; and

2) associated algorithms for recognizing these motion concepts.

These algorithms implement a form of feedback by allowing competition and cooperation among local hypotheses. They also allow a change of attention mechanism that is based on similarity links between knowledge units, and a hypothesis ranking scheme based on updating of certainty factors that reflect the hypothesis set inertia.

The framework is being realized with a system called ALVEN. The purpose behind this system is to provide an evolving research prototype for experimenting with the analysis of certain classes of biomedical imagery, and for refining and quantifying the body of relevant medical knowledge.

*Index Terms*—Artificial intelligence, computer vision, knowledge-based systems, left ventricular wall motion analysis, motion understanding.

## I. Introduction

### *Motivation*

A FRAMEWORK has been developed for the abstraction of motion concepts from sequences of images by computer. This framework is being tested through the implementation of a system called ALVEN (A Left Ventricular Wall Motion Analysis Consultant) that analyzes films of the human left ventricle and generates a conceptual description of the shapes and motions exhibited by the left ventricular wall, noting abnormalities and unusual occurrences. A complete current description of ALVEN can be found in [9].

From an artificial intelligence point of view, the research conducted in the ALVEN project is concerned with the problems inherent in designing computer-based "motion understanding" systems. In our view, this research has the potential of contributing a new framework for research on computer vision which, unlike earlier ones, accounts fully for the "pragmatics" of visual understanding (i.e., the context within which each incoming picture should be analyzed). Clearly, such a framework will not always be applicable. We claim, however, that there are many domains for which representations are most naturally structured around motion concepts, and in which motion information provides a useful guiding principle for the analysis of images. The importance of this research for expert computer vision systems rests on this claim.

The framework to be described is based on the assumption that the design of an expert system for the performance of some task requires the representation and use of knowledge relevant to that task. This assumption has forced us to confront the problem of knowledge representation for shape and motion concepts, as well as for concepts related to left ventricular wall motion analysis. Furthermore, since most of this knowledge is verbal and qualitative, we were forced to address the interface between this kind of knowledge base and quantitative image data.

From a medical point of view, ALVEN's importance rests on its ability to assess the human left ventricle (hereafter LV), as a functioning muscle. This assessment must be based on the velocity profile of the LV wall segments. This calculation of such profiles and their descriptions in qualitative terms is a problem that has not yet been tackled successfully by medical researchers, although some progress has been made towards its solution.

The visual data for assessing LV wall motion are provided by X-ray cineangiography, with an image rate of 60 images/s. When the heart muscle is impaired or damaged, abnormalities occur regionally, and different segments of the LV often behave differently. Furthermore, there is some evidence that if one segment is damaged, another may "overperform" to make up this deficit on a more global scale. Thus, if the ventricle is to be assessed as a muscle, it must be assessed regionally, as well as globally. The extent, velocity and acceleration of contraction must be measured for as many regions as are necessary to determine muscle competence.

There are two central difficulties facing any approach to the analysis of cinecardioangiograms: the huge number of images that must be analyzed, and the poor quality of the individual images due to X-ray dosage limitations. These difficulties make an objective and consistent analysis very difficult for humans, and provide the motivation for selecting this problem domain as the application area for our research. From a medical point of view, the aim of the research is twofold:

1) to produce a system that can aid in the qualitative and quantitative analysis of cinecardioangiograms, and

2) to provide a framework for the representation and quantification of medical knowledge about the dynamics of the human left ventricle.

ALVEN is a first step towards achieving these objectives.

Manuscript received January 25, 1980; revised April 3, 1980.

J. K. Tsotsos and H. D. Covvey are with the Cardiovascular Unit, Toronto General Hospital, Toronto, Ont., Canada, and the Department of Computer Science, University of Toronto, Toronto, Ont., Canada.

J. Mylopoulos is with the Department of Computer Science, University of Toronto, Toronto, Ont., Canada.

S. W. Zucker is with the Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada.

0162-8828/80/1100-0563$00.75 

*What is a Motion Understanding System?*

Let us first define what we mean by the term "motion understanding" (MU). It is not enough to just determine the movements or changes that occur between a pair of consecutive images ("interimage descriptions" [7]). In many cases it is possible (indeed, it is desirable) to use a single motion concept to describe the changes exhibited in several consecutive images. Such a motion term would abstract a summary description from the multitude of interimage descriptions generated for the sequence. As a simple example, consider a sequence of images that show a person walking. If only interimage descriptions are used, the result is a complicated and cumbersome set of terms. It is not only desirable, but it would be necessary if we wish to claim that our system has "understood" these movements, to summarize them as a "walk" motion. Therefore, a system that can provide interimage descriptions as well as "sequence spanning" descriptions, with the semantic components of the specific motion terms used being stored in a knowledge base that is used to guide the recognition process, will be termed a *motion understanding system.*

The problem of motion understanding can be broken down into several subproblems.

1) Computer vision:
   a) image segmentation and object recognition
   b) object description
   c) motion detection
   d) motion tracking
   e) interimage movement description.
2) Representation of knowledge:
   a) general temporal concept representation
   b) problem domain motion concept representation
   c) recognition biased knowledge organization.
3) Recognition control structure:
   a) integration of descriptive and visual concepts
   b) change and focus of attention mechanisms
   c) temporal segmentation
   d) disambiguation due to object occlusion
   e) goodness-of-fit measures
   f) generation of low-level guidance
   g) scene sampling rate considerations
   h) artifactual motion handling, i.e., "temporal noise"
   i) generation of sequence spanning descriptions.

In the main, past motion research has dealt with the problems under the first heading above—the vision aspects of motion [4], [7] with very little work in the last two main subtasks. Our research follows generally the approaches of past research in the vision aspects, but expands upon the representation and control aspects. Specifically, we address each of the subtasks outlined above under the representation and control headings with the exception of disambiguation due to object occlusion. In addition, we assume that all the motions are two-dimensional (no depth information), that the observer is fixed and plays no role in the descriptive process, and that causal information and descriptions are not part of the system's repertoire. In order to simplify the process so that we concentrate mainly on the study of motion, we assume that a conceptual description of the first image of the sequence is provided.

## II. Representation of Motion Concepts

*Overview*

Our representation has its roots in semantic network theory [2], and in particular, in the PSN formalism [3]. Each motion concept has a ***frame*** associated with it that defines it. A frame is not a fixed unit: it is definable by the user and may encompass as large a knowledge unit as desired. Frames have an arbitrary number of ***slots*** that form their parts, and each slot has an associated ***type*** that refers to another frame, thus defining a ***PART_OF*** hierarchy of description. Each level down in this hierarchy provides a more detailed form of description for the motion concept, spanning all the levels between the most abstract motion terms to the picture elements in the images. Slots come in two varieties: ***prerequisite*** and ***dependent.*** Prerequisite slots specify concepts that must be observed before the frame can be instantiated, while dependents provide additional semantic components that are included along with the frame concept on instantiation. In addition, slots can have constraints and defaults associated with them, and associated with each of those is an ***exception frame*** that is generated when the constraint is violated by the input data.

A simple example of a frame is given below. It defines the knowledge the system uses to recognize the concept of contraction. It is defined in terms of the primitive frame AREA_CHANGE.

```
frame AREA_CHANGE with
    prerequisites
        subj: CONTRACTILE_OBJECT;
        time_int: TIME_INTERVAL:
        start_a: AREA_VALUE;
        end_a: AREA_VALUE;
end
frame CONTRACT is_a AREA_CHANGE with
    prerequisites
        start_a: such that
            start_a > end_a;
    dependents
        speed: SPEED_V with
            speed ← (start_a-end_a) ÷ time_int.duration;
end
```

These two frames illustrate most of the syntactical constructs of our representation formalism. The "speed" of contraction is defined as the rate of contraction. It is determined from the initial and final areas of the contractile object exhibiting the change in area, divided by the duration of the change. The "duration" refers to the slot of the TIME_INTERVAL token associated with the slot "time_int" in the AREA_CHANGE frame.

Suppose that we wish to define a more specific type of contraction, one which, for example, has a specific rate of contraction. Such a constraint could be added to the "speed" slot so that it becomes

```
speed: SPEED_V with
    speed ← (start_a-end_a) ÷ time_int.duration
    such that (speed > 5 and speed < 12)
    exception TOO_SLOW with
        subj ← self.subj
        time_int ← self.time_int;
```

Here we have also included the exception "TOO_SLOW" which would be instantiated if the observed rate of contraction were slower than that specified by the constraint.

The two frames above are organized in terms of the *IS_A* relationship. This relationship organizes the frames into a second hierarchy, the IS_A hierarchy, which places more general frames at the top and more specialized ones at the bottom. IS_A implies inheritance of properties from father to son frame in the knowledge base. A son frame can refer to any slot of a father frame simply by the slot name. We have observed that the IS_A and PART_OF relationships are sufficient for automatic creation of sequence spanning descriptions in certain cases as a result of the instantiation process. (See [9] for further discussion.)

***Similarity links*** [6] relate one frame to another and aid in altering the system's active hypothesis set. Similarity links include information such as common components of two frames, the time course of differences that must be observed in the image sequence in order to discriminate between the two frames, and activation binding information. Using such a formalism, a knowledge base is defined for the motion concepts necessary for the application domain.

Specific frame instances are linked to the frames with the *INSTANCE_OF* relationship. This relationship compels an instance frame or ***token*** to have the structure dictated by its type, i.e., the frame it is associated with through the INSTANCE_OF relationship. Objects bound to slots in a token are determined either from the image (prerequisites) or through calculations (dependents). More details about the knowledge representation formalism can be found in [9].

### *Motion Concepts*

In natural language, the tools used for description of motion concepts are verbs such as "cross," "lift," and "run"; directional prepositions such as "towards," "through," and "over"; and adverbs such as "slowly" or "quickly." In order for these concepts to make sense in a sentence, certain semantic components must be present. These components completely define the motion concept. Miller [5] has done a very good analysis of the semantic components for a large class of English motion verbs. Using Miller's work as a foundation, Badler [1] refined these concepts, gave detailed definitions of the directionals, and outlined a framework within which recognition can be done. Further investigation of the semantic components of motion verbs and on their representation appears in [8]. The work presented here is strongly based on these three research efforts.

Miller did his analysis from a linguistic point of view, using natural language. Components such as "propellant" or "permissive," which describe the intent of the agent of the action, may be impossible to detect from a sequence of images. "Speaker" is nonexistent but may be interpreted as "observer," and we assumed that the observer plays no role in the motion description process. In addition, there are no components for the description of the changes in an object's physical properties such as "expansion" or "joining two objects to form one." The components only describe motions of rigid or jointed objects. We will not include three of Miller's components: causative, instrumental, and deictic. The first two require causative knowledge which is not part of ALVEN. The latter requires participation of the observer in the description process as well as three-dimensional data, and these are also not part of ALVEN.

Let us now discuss the motion semantic components that are included in ALVEN, defining them in a way that is appropriate for use in a visual motion understanding system.

*Motion:* An object, of any type, has exhibited a motion if a change of location has been observed for its centroid, or for any of its parts. This definition may be applied recursively to any of the object's parts.

*Reflexive-Objective:* The distinction between reflexive and objective verbs or directionals depends on the reference point for the descriptive term used. If the reference is the computer imposed coordinate axes, then the descriptive term to be used is reflexive. Examples of reflexive terms are the directionals "upwards" and "leftwards" or the verbs "rise" and "fall." If the description is in terms of some other object, whether it is in motion or not, then the term to be used is objective. Examples of objective directional prepositions are "towards" or "over," while verbs are "cross" or "approach."

*Permissive:* This component implies knowledge of the intent of the agent exhibiting the motion. However, several verbs such as "drops" or "releases" do not require intent in order to be recognized. Such motions require complex semantics: the agent of the action must initially be restraining the object of the motion so that the object cannot move in some specific direction or with some specific velocity. The agent then exhibits some motion, which is dependent on the agent's type, that frees the object to exhibit some motion that it previously could not exhibit.

*Propellant:* The application of force can be considered for a vision system so that agent intent or forces such as gravity or magnetism need not be considered. The participating objects must be in contact, at least one of them must be capable of self-propulsion (have a motion capability of "motile"), and the resulting motion of the nonmotile object is in the same direction as the motion of the motile object that applies the force. Examples of such verbs are "lift," "lower," "push," "pull," "shove," or "drag."

*Directional:* Directionals, as mentioned previously, have been adequately defined by Badler [1]; his definitions will be used with little or no modification. Examples of directionals are "across," "through," "into," etc. In general, all objective directional definitions involve a sequence of changes in the spatial relations between the participants.

*Medium:* The medium here defines the element in which the motion is taking place whether it be on land, in the air, in the water or in a patient's chest. For our application domain, it is

clear that the medium is a patient's chest. Therefore, the medium will not be explicitly referred to, but the methodology could take care of it if the problem domain required it.

*Inchoative:* This defines changes of an object's condition or spatial relations. An example is "the door opened"—rephrased so that the inchoative component becomes more clear, "the door became open." The change described by the verb "open" is that the size of the aperture between the door proper and the door jamb becomes larger. The form of the inchoative component is generally "becomes adjective," and in many cases, single verbs may be found that describe the change. Such changes in physical properties or spatial relations are not well defined by Miller and are expanded upon by ALVEN. ALVEN contains several physical property change concepts including changes of length and of area. See [9] for further discussion.

*Change-of-Motion:* This component describes not only the beginning or the ending of a motion, but also the change from one motion class to another for the same object when the two motions are adjacent in time. This component appears only implicitly as the start and end time relationships for any motion and the constraints on which motions may preceed or follow it. Instances of this component are determined by the temporal segmentation process.

*Velocity:* Velocity terms such as "slowly" or "quickly" are used in relation to the object's motion capabilities. For example, a car moving at 30 km/h is moving slowly, while that speed for a person is quite fast. Thus, if velocity descriptions are required, they must be defined individually for every class of objects in the system's domain.

Our knowledge representation was designed to accommodate all of the above motion semantic components, except for those involving causation or three-dimensional movements. These components appear in our knowledge base in diverse ways. If an object exhibits a ***motion*** then some frame within the motion hierarchy is instantiated to represent the motion, with the ***subject*** of the motion being the object which exhibits the motion. Subjects will always appear under the slot name "subj". In addition to the "subj" slot, there is always a time interval during which the motion occurs. This slot is always named "time_int". ***Reflexive*** and ***objective*** motions are distinguished by the absence or presence respectively of a ***referent*** for the motion. The "ref" slot denotes the object to which the motion is compared. ***Permissive*** or ***propellant*** components do not appear explicitly, but simply define different classes of example verbs. Such verbs have both an ***agent*** which appears under the "agent" slot and represents the object which is producing the motions, and an ***object*** which appears under the "obj" slot and represents the object which is being affected by the motion of the agent. ***Directionals*** are used in two ways—numerical trajectories or qualitative directional terms. If a numerical ***trajectory*** is used to define a motion, it appears under the slot name "traj", while directionals always use the slot name "dir". ***Velocity*** is included as the slot "speed". ***Inchoative*** components again are not included explicitly, but provide a different class of verbs. These verbs are all characterized by the presence of two slots, the start and end states, and these slots refer to the subject of the frame, which is always a physical property of the object (such as "length" or "perimeter" or "width" or "shape"). The ***change-of-motion*** component, as explained above, appears as a timing constraint. If, for example, one motion *A* must immediately precede another *B*, then the constraint is

A.time_int.end_time = B.time_int.start_time

All motions have "start_time" and "end_time" slots and thus for any motion the frame may define the relative start and end time of the motion and any constraints on which motions may preceed or follow it.

### *Primitive Interimage Description Terms*

A small set of primitive movement descriptors has been determined in terms of which all higher level motion concepts are described (using a PART_OF hierarchy). These descriptors encompass all the basic concepts of the motion semantic components described above. They are TIME_INTERVAL, LOCATION_CHANGE, LENGTH_CHANGE, AREA_CHANGE, and SHAPE_CHANGE. Each of them has its semantics defined by a frame. The TIME_INTERVAL frame is instantiated for each interimage interval, and, since each motion concept has "time_int" as a slot, all change of motion semantics can be expressed via the TIME_INTERVAL tokens. All motions involve some type of LOCATION_CHANGE, whether it be at the picture element level or at the object level. In addition, all changes of spatial relations between objects involve location changes, as well as all trajectory, direction, and speed information. Inchoative components are handled by the final three primitives: SHAPE_CHANGE, LENGTH_CHANGE, and AREA_CHANGE. These are also particularly important for our application domain of left ventricular wall motion.

Below, we show the LOCATION_CHANGE frame and also a specialization of it, the TRANSLATE frame, as well as a further specialization, the RIGHTWARDS frame. This illustrates both the use of the PART_OF and IS_A relationships. Many more examples, as well as the complete frames for normal and abnormal heart cycles appear in [9].

```
frame LOCATION_CHANGE with
    prerequisites
        subj: POINT such that
            for ob1: OBJECT where
                ob1.mot_cap = mobile
                subj part_of ob1;
        time_int: TIME_INTERVAL;
        start_loc: POINT;
        end_loc: POINT;
end
frame TRANSLATE is_a LOCATION_CHANGE with
    prerequisites
        subj: such that subj instance_of CENTROID;
    dependents
        subj: OBJECT with subj ← ob1;
        speed: SPEED_V with
```

```
                    speed ← (((end_loc.y - start_loc.y) ** 2 +
                            (end_loc.x - start_loc.x) ** 2) ** 0.5) ÷ time_int.duration;
                traj: TRAJ_V with
                    traj ← arctan((end_loc.y - start_loc.y) ÷
                                (end_loc.x - start_loc.x));
end
frame RIGHTWARDS is_a REFLEXIVE_TRANSLATE with
    prerequisites
        traj: such that
            (traj ≥ 0 and traj < 3*π/8) or
            (traj ≤ 2*π and traj > 13*π/8);
end
```

*Organization of Motion Concepts*

The motion concepts described thus far can be organized using the IS_A relationship as noted above. We have defined a hierarchy of motions that specifies the structure and semantics that any problem-domain concepts must conform to. One side of the hierarchy describes motion concepts relevant to single object movements with respect to a fixed point of reference (e.g., object rotation). Single objects are assumed to be indivisible and may represent an entity or part of an entity. Such simple motions are further classified into changes of location (LOCATION_CHANGE) or changes of physical properties (PHYS_PROP_CHANGE). In turn, LOCATION_CHANGE's are either ROTATE's or TRANSLATE's, while PHYS_PROP_CHANGES's are either AREA_CHANGE's, LENGTH_CHANGE's, or SHAPE_CHANGE's. Examples of TRANSLATE's are verbs such as "fall" or "approach"; examples of AREA_CHANGE's are "contract" or "expand"; examples of LENGTH_CHANGE's are "widen" and "compress." The other side of the hierarchy deals with aggregate motions, i.e., motions of the following kinds: motions of objects defined in terms of the motions of their parts, (SIMUL_MOT_PARTS); superimposed motions for a single object, (SIMUL_MOT); motions involving independent, simultaneously moving participants, (SIMUL_DISTINCT_PARTS); and, sequences of motions for a single object (SEQUENCE). Examples of SIMUL_MOT_PARTS motions are "forward right leg swing" or "left ventricular segmental contraction"; examples of SIMUL_MOT motions are verbs such as "walk" or "run"; examples of SEQUENCE's are "walk in place" or "heart beat"; and, examples of SIMUL_DISTINCT_PARTS are verbs such as "lift," "drop," "pull," or "release." This classification of motion concepts is based on work described in [8]. Examples of the actual frames for each class and of sample verbs can be found in [9].

## III. System Overview

The paradigms of competition and cooperation among hypotheses and hypothesize-and-test form the basis of our recognition control structure. The key feature of the control structure is that it is driven by the organization of the knowledge base, that is, by the primitive relations between knowledge units. A feedback loop is incorporated in order to link the several components of the structure.

Processing proceeds image by image. The system looks at one image at a time, and once it is finished, the image is discarded. A simplificiation for our purposes is that the objects in the first image in the sequence be identified and classified prior to motion analysis. Expectations produced by the system guide the low level process in identifying objects in subsequent images. Expectations are in the form of region and orientation biases for a modified relaxation labeling process [10]. Such a process only considers pixels within the predicted region and biases against edge elements that are inconsistent with those expected for the region. If an object is not found at precisely the same location in the image as predicted by the high levels of the system, then the object has moved and presumably, motion has been detected. Other portions of the system are responsible for weeding out true motions from artifactual ones. This motion is one of two forms: either the object is being tracked, or the object has just started or stopped moving. If the object is being tracked, the predictions specify a range of possible locations for it in the next image. The object descriptions are in terms of the low level vision, constructs, such as edges in the image, along with descriptors such as type, axes, area, and arc length. This description is called the ***essential trace*** for the object.

Pairs of timewise adjacent essential traces are combined into the ***essential kineses*** of each object. These are determined using a set of matching heuristics such as similarity of shape type and of location. Essential kineses are defined in terms of location changes of points, length changes of axes and perimeters, area changes, and shape changes. These four primitive kineses provide an intermediate representation for relating quantitative changes to qualitative ones. The kineses are used to match against the hypotheses which are active for a particular object's motion.

The system must start with an initial hypothesis for the motion of each object in the image sequence. This initial "guess" need not be a perfect one; however, it is assumed that the knowledge base's hypotheses are properly related to one another and that it is complete in this respect. The essential kineses are described in the same terms as the lowest level of motion description for each hypothesis of the knowledge base. It may be necessary in some cases (when the motions are only very coarsely described) to abstract higher level descriptive terms from the kineses. Matching failures between expected and actual kineses are used by the change of attention mechanism. These failures are represented in terms of ***exception***

*frames* which contain any information necessary to the change of attention process so that proper selection can be made of alternate hypotheses. This selection is made via the similarity links which are present in the hypotheses. The time course of differences consist of exception tokens, while the time intervals specified in each token give timing information. Each active hypothesis is related to the other active ones by its ***conceptual adjacency***. Conceptual adjacency is defined in terms of the following primitive relationships of the knowledge base: IS_A, PART_OF, similarity, and time course.

Since several simultaneous hypotheses can coexist, a focus of attention mechanism is necessary in order to limit the number of hypotheses under consideration. Our focus of attention mechanism uses a property of feedback systems: inertia. That is, the output of a feedback system changes slowly over time and therefore changes in focus are continuous. We do not wish to be faced with the problem of erratically shifting foci. The system focuses on the best hypotheses under consideration using certainty factors that are attached to each hypothesis. The certainty factors are updated using relaxation labeling [10] with dynamic neighbourhoods and compatibilities that are determined using the conceptual adjacency between hypotheses and hypothesis matching progress. A single iteration of relaxation is applied between images in order to preserve the inertia of the feedback system. Iteration is not done until convergence to stable certainty factors is achieved.

The problems encountered by other large application systems are due to inexact information—either incorrect data, extraneous information (part of the problem domain but not directly related to the task at hand), or noisy data. This causes mismatching of hypotheses, leading to erroneous partial matches and multiple matches where in reality only one hypothesis should match. Consider these problems in the simpler context of edge detection in single images. Line quantization, lighting, and shading effects cause erroneous input, while noise adds another dimension of error to the process. With feedback, as in relaxation labeling [10], proper functioning under such conditions is possible. The reason is that isolated instances of stimuli are rejected. For example, in static situations, such as in a single image, a noise point is not related in a meaningful way to its neighbors in the image. Similarly, in dynamic situations, where events change with time, spurious events are similarly unrelated to the events in a sequence or context, i.e., its neighbors in time. The relaxation procedure, by using slowly responding feedback, can reduce the effect of such erroneous situations.

The response of a focus mechanism based on feedback is not instantaneous—there is an inherent delay. Past systems have attempted to provide a completely updated view of the analysis of the problem domain at each instant during processing. Therefore, any errors in the input data would be completely integrated as if they were valid data, rather than waiting to see if the trend is indeed that which is suggested by the data. The focus of the system should not shift abruptly—it should only change if the stimulus dictating the change is present over a significant period of time. Such time periods are clearly problem dependent.

In motion understanding, such a concept is vital. Sequences of images present difficult problems: occlusion of objects, so that locations may sometimes only be guessed at; noisy individual images within the sequence; hypotheses with overlapping expected time intervals which describe the possible motion concepts of the domain; and, particularly for our domain of left ventricular wall motion, poor contrast images. Feedback principles allow us to tackle these problems. One way in which isolated stimuli such as noise or errors can be handled is by providing descriptions in the KB at varying levels of abstraction, from coarse to detailed. Feedback can then be present between these levels, with the coarse description placing strong, more global constraints on the detailed description, thus assisting in the removal of isolated error stimuli. In addition, since the relaxation process accumulates the evidence of the matching history of a hypothesis, it is very difficult for an isolated error event to negate a large positive history of successful matching (or, for that matter, a history of matching failures).

The focus of attention mechanism ranks the hypotheses on the basis of their updated certainty factors in order to determine the best ones. Each hypothesis, when activated, receives an initial certainty factor equal to that of the hypothesis that activates it. A relaxation process is then used to update the certainty factors. The relaxation process is based on conceptual adjacency that specifies which hypotheses are competitors and which ones are complementary and in what respect. The best hypothesis (highest ranked) then are used to derive the expectations for the next image.

The instantiation of a hypothesis poses a unique problem: the ***temporal segmentation*** of timewise adjacent events. In other words, where does one motion start and the previous one end for a particular object? This problem is solved by analysis of the time course of certainty factors for the two motions involved, and is described in a subsequent section.

### *Hypothesis Organization and Change of Attention*

Active hypotheses are organized by their "conceptual adjacency." If two hypotheses define motions for the same subject and for the same time interval, then we say that they satisfy the ***common subject-time*** (CST) condition. This concept is used in the following discussion. By definition, two hypotheses H and H′ are adjacent conceptually, if one (or more) of the following hold.

1) There is an active similarity link between H and H′, and H and H′ satisfy the CST condition. Such an adjacency implies mutual exclusion, i.e., only one of the hypotheses can be instantiated.

2) H precedes H′ in time or H follows H′ in time, and they define motions for the same motion subjects. In this case, there is overlap of time interval and since only one motion can exist during any interval, the hypotheses are competitors. The same problem exists in speech understanding systems at word boundaries.

3) H IS_A H′ and CST holds.

4) H has H′ as a direct IS_A descendent, and the CST condition holds.

5) H is adjacent to itself for the next time frame. This handles exceptions in the parts of hypothesis H over time. Thus, the PART_OF hierarchy is implicitly taken into account.

The system adds hypotheses to its list of active ones via activated similarity links and via the "next" temporal constraint, (i.e., if motion X is the "next" one after Y, motion Y has an end time which is the same as motion X's start time).

A similarity link in hypothesis H is activated between hypotheses H and H′ when:

1) the similarity expression is satisfied, i.e., all properties of H′ which must be true before H′ can be activated must have been instantiated; and,

2) at least one of the exceptions in the difference expression has been instantiated. It is not necessary that this be the first one in the time course expression since noise effects may mask it. One must realize, however, that noise or extraneous information may erroneously trigger the similarity.

There are several additional considerations that arise when determining similarities between hypotheses.

1) When a particular exception cannot be handled by the local similarity links, the links of its IS_A ancestors are checked.

2) Since activation of a frame automatically activates its IS_A ancestors, transfer of parts between the two frames' IS_A ancestors may be accomplished by the binding expressions of the similarity links (if present) between those frames.

3) There is also a need to propagate similarities upwards along the PART_OF hierarchy, because possible mismatches of a motion component may require completely new contexts for the newly activated parts. This is done by automatically raising an exception in the parent hypothesis. When an exception is detected in a part A of frame B, this automatically generates an additional exception for frame B stating that frame B has failed. In this way exceptions are propagated up the PART_OF hierarchy. The exception carries with it a special slot "prereq_part" which specifies the newly activated frame name C. The parent frame of B, that is D, can use its similarity links with the added constraint that C must be PART_OF any newly activated destination frame of D.

The start time of the activated hypothesis is taken to be the instant of activation of the activating hypothesis. This is reasonable because if it were true that the activated hypothesis should have started earlier, then a similarity link from a previous hypothesis in time for that object should have activated it. If it should start later, then the two hypotheses should be related by the "next" time constraint. Such considerations are true only if both of the hypotheses specify durations or start and end times. If they do not, then it is ambiguous as to whether the motions follow one another or are competitors for the same time interval. If a hypothesis A activates B through a similarity link and no timing information is given, then this is considered to be a "next" constraint. The description produced would be that the object exhibits motion A for a duration of X and then motion B for a duration of Y.

Suppose a hypothesis A activates two or more hypotheses via similarity links at the same time instant. Also, assume that there is no time information specified. Then frame A has a "next" relation with each of the newly activated ones, while the new ones are in competition with one another for that time interval. If time information is provided, then the frames are all similar to one another.

*Hypothesis Rating–Focusing in on the Best Hypotheses*

A focus of attention mechanism is needed because the prediction mechanism must know which are the "best" hypotheses at any instant in the processing, so that it can base its expectations on them. The hypotheses will be ordered by means of the certainty that the system has in them. The best candidates are determined for each object; that is, hypotheses certainty values are not compared unless the hypotheses define motions for the same object and for the same time interval, (the common subject-time condition). Thus, each object would have a leading hypothesis, and each part of each object would have a leading hypothesis. It is not necessarily true that the best hypothesis for an object's part is PART_OF related to the best hypothesis for the object as a whole.

Each hypothesis has a certainty factor associated with it–a number between 0.0 (completely uncertain) and 1.0 (completely certain). Initial values are set through the conceptual adjacencies of each hypothesis when it is activated. If a hypothesis H has N competitors (either mutually exclusive ones, through similarity links, or through the "next" time constraint), then each hypothesis of the competing set has an initial certainty of $1/N * \text{cert}(H)$. If a hypothesis has no competitors, then as far as the system is concerned, it is certain that the hypothesis will be instantiated–because there is no other possibility. This is why similarity links are so crucial–matching failures would not be recorded in a hypothesis' certainty factor through the relaxation process unless they also activate a similarity link to a competing frame. The reason for this is that certainty factors are normalized over the set of competing hypotheses. In relaxation labeling for edge detection, the competing labels for a point were normalized so that the sum of their certainties remained at 1.0. Here as well, the competing labels for any object's motion are normalized. When a hypothesis activates a similar frame, they share equally the activating hypothesis' certainty factor.

Once the initial certainties are set for a hypothesis, the certainty is updated for each subsequent image. The updating rule that we use is based on relaxation labeling [10]. The only syntactic change is that we need to keep track of the certainty factors at each time instant $t$. Its form is the following:

$$cert(t+1, H) = \frac{cert(t, H) * [1 + q(H)]}{norm\ (H)}$$

where $cert(t+1, H)$ is the certainty factor for hypothesis H at time instant, $t+1$, $q(H)$ is the contribution from its neighboring hypotheses (assume that there are $n$ of them), and is given by

$$q(H) = \sum_{i=1}^{n} w(H, H_i, t) * comp(H, H_i, t) * cert(t, H_i).$$

$norm(H)$ is the normalizing factor for hypothesis $H$. It is

given by

$$norm(H) = \sum_{i=1}^{n} cert(t, H_i) * [1 + q(H_i)]$$

and $w$ is the weighting function of the contribution from one hypothesis to its neighbor at a particular time instant. The sum of the contributions must be unity. Finally, *comp* is the compatibility function between two hypotheses at a particular time instant.

The compatibility value is dynamic—it is determined by a function relating matching progress in the current image for hypothesis $H_i$, the current time instant and the conceptual adjacency type exhibited by $H$ and $H_i$. If $H_i$ fails to match in the current image, it lends support to $H$ (positive compatibility), while otherwise it removes support, (negative compatibility). Rules for selecting values for the various types of compatibilities can be found in [9].

The major change from the process described by Zucker *et al.* is that we do not iterate until certainty factors converge to stable values. Rather, certainties are updated using only one application of the rule. In this way, each update corresponds to an analyzed image in the sequence. The reason for this change requires a clear understanding of what the relaxation labeling process (RLP) does.

For edge determination in single images, the responses of an edge operator supply the initial certainty factors for the possible labels for each point. For the purposes of this argument, let us assume that there are three labels. This initial assignment of certainty factors place the starting point of the RLP at some point, say A, on the plane formed by the certainty factor vectors {1, 0, 0}, {0, 1, 0}, {0, 0, 1}. The plane is as it is because the sum of the certainty factors must remain at one; otherwise, the process is not gauranteed to converge. From point A, the RLP, using good edge continuity criteria takes the certainty factor vector through some path to a final convergent labeling. Suppose that in our motion case, the same were done between images. The effect is disastrous. Because of the nature of the RLP, the system could never move away from this final stable state, and therefore after the first pair of images, the certainty factors could not change. This is clearly undesirable.

Using the concept of hypothesis set inertia, we can now justify the design of the updating mechanism. If only one application of the updating rule is used between images, the effect is that the system moves along a short vector tangent to the path that it would have taken if many iterations were done. Two images are enough to determine this vector. The system thus has inertia of location because it does not move very far away from its current location. This is desirable because hypotheses and their compatibility values are uncertain. The length of the vector is determined by how consistent the matchings of hypotheses and the kineses are with the constraints imposed by the organizational axes of the KB. The system also has inertia of direction because it moves along a tangent to its current path in the plane formed by the hypothesis space. The problem of rapidly shifting foci is therefore greatly reduced. However, other considerations arise. What

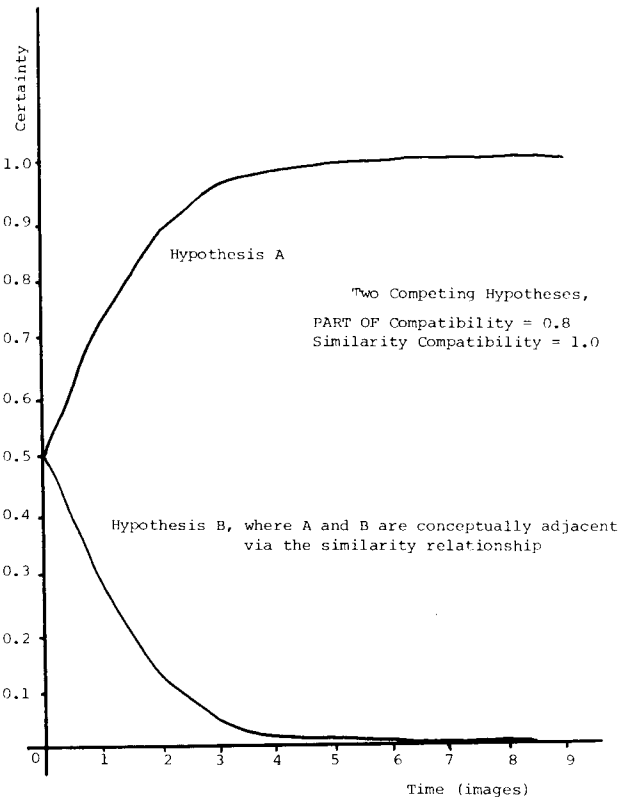



Fig. 1.

are the limitations of such a process in terms of how many iterations (in other words, images) are needed in order to determine a consistent labeling, and how is such a process affected by extraneous stimuli? Experimental evidence presented below answers these questions and provides support for our design decisions.

Fig. 1 shows two similar hypotheses competing for description of a common time interval and demonstrates how consistent matching history leads to the selection of the correct hypothesis. Note that the curves look very much like exponential functions. In electronics, when the response of a circuit exhibits such a shape, a useful parameter which is calculated is the system "time constant." In particular, if the curve can be approximated by the function

$$V = V_0(1 - e^{-(t/\tau)})$$

for the rising curve, where $V$ is the system output and $V_0$ is the final, stable system output, then $\tau$ is called the "time constant" of the system. The output comes to within $1/e$ of its final value in a time interval equal to $\tau$. Or, in other words, the certainty reaches 0.632 if 1.0 is to be its final value. We can determine values for $\tau$ in our system as well, with an analogous interpretation. If we use the definition of the critical value given above, we can say that the system must analyze $\tau$ images in order to discriminate between two competing hypotheses under ideal matching conditions. In the example given, this response time is three images. However, we must be aware of the fact that the similarity competition is not the only rating mechanism operating in this example. There is also the PART_OF compatibility to take into consideration.

After trying all possible combinations of compatibilities, we observe that the no combination leads to a smaller $\tau$ than three images.

The PART_OF compatibility can be viewed as a "noise sensitivity" parameter. The greater the probability of erroneous essential kineses, then the smaller the PART_OF compatibility should be. However, the price that is paid is that the final values are not very far apart. That is, there is a low "selectivity" among the hypotheses. However, we do observe that the final ordering of hypotheses is independent of the specific values of compatibilities used, again verifying the proper functioning of this mechanism.

The graph in Fig. 2 shows the results of tests of the certainty updating mechanism under the following conditions.

1) Each point is the average time value obtained over 30 trials at the same values for similarity and PART_OF compatibility and number of competitors.

2) The trials are run using a sequence of random numbers to represent matching successes and failures, and are adjusted for noise levels.

3) The decision threshold, that is, the certainty value that must be achieved before a particular hypothesis is instantiated is related to the number of hypotheses in the following manner. The $\tau_i$ is the time at which the certainty of a hypothesis reaches

$$\left(1 - \frac{1}{N_i}\right) * (1 - e^{-1}) + \frac{1}{N_i}$$

which simplifies to

$$0.632 + \frac{0.368}{N_i}.$$

This is determined using the assumption that each of the hypotheses that are competing start with equal certainty values, determined using the number of hypotheses $N_i$.

1) Similarity compatibility is set to 1.0 in all cases.

2) Noise-free and 10, 20, 30, 40, and 50 percent noise images were tested. For our test, 10 percent noise implies that one out of ten data items passed to the hypothesis matching process will not match the expected data ranges when in reality, the image data will indeed match.

3) Each test was run on from 2-16 simultaneous competitors.

The results are approximated by linear relationships, although there is no reason why we should believe that the relationships should be linear. The 50 percent noise line is very close to vertical--thus confirming the expectation we get from information theory that at 50 percent noise there is no significant data in the images. The remainder of the lines verify our intuition in expecting that the noisier the data, the more of it that is required in order to make decisions with the same confidence.

Another interesting observation can be made from this graph. If there are, for example, 5 competing hypotheses, under 10 percent noise conditions, then the system must examine at least 10 interimage descriptions (11 images) in order to discriminate among them. This means that if the motion concepts represented by the hypotheses have expected minimum duration D, then the sampling rate of the film must be at least 11/D in images/s. This can be generalized to the following relationship:

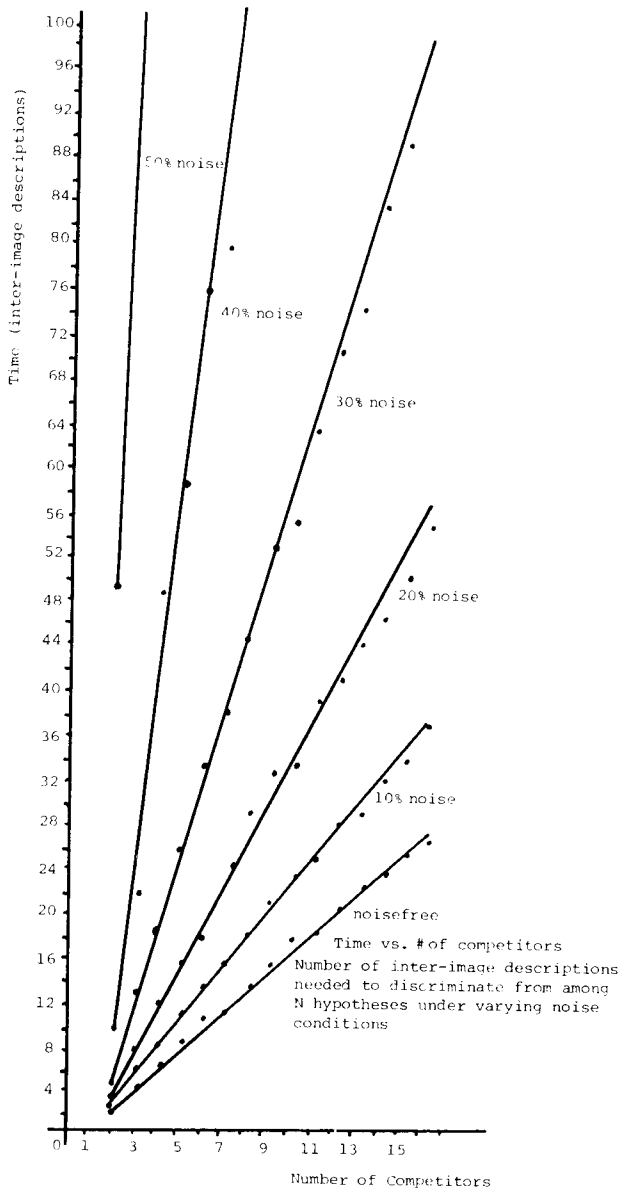



Fig. 2.

$$SR_{\min} = \max_i \frac{(\tau_i + 1)}{dur_i}$$

where $SR_{\min}$ is the minimum required sampling rate, and $\tau_i$ is the minimum number of interimage descriptions required determined from the graph, for discriminating from among a set of hypotheses {H}, cardinality of which is $N_i$ and whose minimum duration is $dur_i$.

This relationship has important implications. It can be used to determine whether or not a particular image sequence related to a specific knowledge base, can be analyzed using the methods presented in this paper.

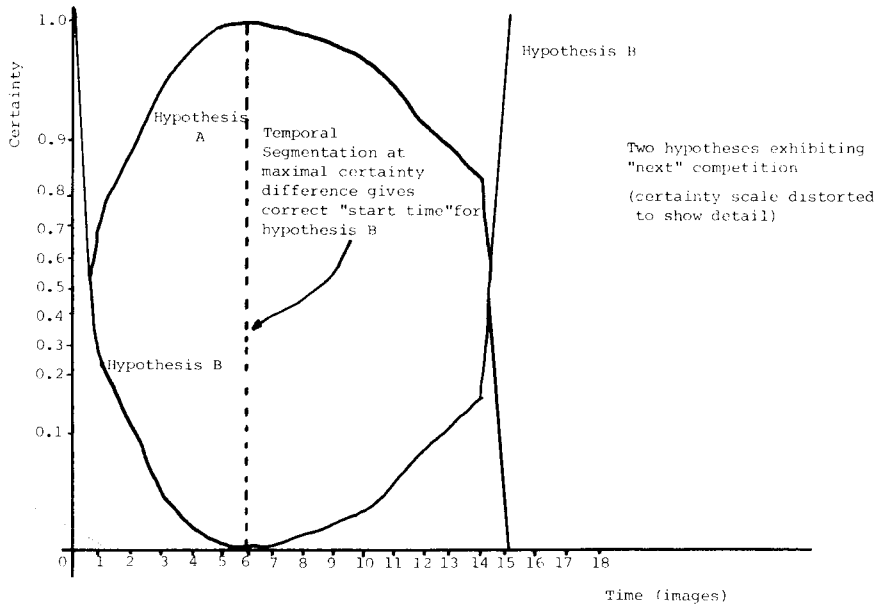



Fig. 3.

*Temporal Segmentation*

Temporal information is a slot in every frame and defines when the motion starts and when it ends. The determination of the start and end times involves both the choice of which is the best hypothesis for an object in a given time interval and which time instants define the interval. We call the determination of this information ***temporal segmentation.***

In our system, the hypotheses considered for temporal segmentation are those in an object's alternative set—the competitors due to active similarity links and those related by the "next" constraint.

Let us consider first the simple case of two hypotheses competing in time, neither of which has any competitors outside the time interval during which they compete. See Fig. 3. Intuitively, we expect the first hypothesis' certainty to be 1.0 until the second one is activated, at which point it abruptly falls to 0.5. At the instant of activation they have equal certainties. Suppose that the motion which corresponds to this second hypothesis does not begin for a couple of images. During those images, the certainty of the first hypothesis will rise, because it is still successfully matching, and that of the second will correspondingly fall. When the second motion begins, the first hypothesis should no longer match. Therefore, its certainty will begin to fall, and the certainty of the second will rise. The point at which the second motion began and the first one ended is marked by the maximum distance between the two certainty values during the interval of competition. This defines their time boundary.

This maximal certainty difference criterion is used for segmentation whenever the "next" constraint is present, regardless of its origin. However, complications may arise when there are several hypotheses competing for the same time interval, i.e., related by the similarity adjacency. If a particular hypothesis H has several possible "next" hypotheses, at the instant H becomes a candidate for instantiation, the best "next" hypothesis is selected (highest certainty), and the segmentation proceeds as above. If there are several simultaneous hypotheses, the best must be chosen. This is done by selecting the one with the highest certainty at the minimum end time of the alternative set. If one waits longer, say until the longest expected duration elapses, then the already ended hypotheses will have their certainty values altered as a result of interactions with their "next" hypotheses. Any "next" hypotheses are then segmented with this best one.

Again, the PART_OF compatibility interacts with the "next" compatibility. In the figure, these values were 0.8 and 1.0, respectively. If we try all possible combinations of values, we observe that no combination leads to a faster overtake than in six images. Therefore, a temporal segmentation procedure is not needed if the time interval represented by six images is insignificant for the error factor in the start and end times of events. In LV motion, this is a significant time interval and therefore the segmentation is necessary.

## IV. Conclusions

A framework for visual motion understanding has been described. The implementation of ALVEN is nearing completion: most components have been tested and several results have been shown. Thus far, although it is clear that much more experimentation is necessary, our results support our design methodology. Complete details of the design, knowledge base for general motion concepts and left ventricular motion concepts, and example operation can found in [9]. Several further projects have been spawned in the medical application area that will further examine applicability of our methodology in other domains: left ventricular function diagnosis, electrocardiogram analysis, and continuous signal understanding.

## Acknowledgment

We wish to thank S. Hume, J. Delgrande, and S. Ho-Tai for programming assistance.

## REFERENCES


[1] N. Badler, "Temporal scene analysis: Conceptual descriptions of object movements," Dep. Comput. Sci., Univ. Toronto, Rep. TR-80, 1975.

[2] R. Brachman, "On the epistemological status of semantic networks in *Associative Networks*, Findler, Ed. New York: Academic, 1979.

[3] H. Levesque and J. Mylopoulos, "A procedural semantics for semantic networks," in *Associative Networks*, Findler, Ed. New York: Academic, 1979.

[4] W. Martin and J. Aggarwal, "SURVEY: Dynamic scene analysis," *Comput. Graphics Image Processing*, vol. 7, 1978.

[5] G. A. Miller, "English verbs of motion: A case study in semantics and lexical memory," in *Coding Processes in Human Memory*, Martin and Melton, Eds. Washington, DC: Winston, 1972.

[6] M. Minsky, "A framework for representing knowledge," in *The Psychology of Computer Vision*, Winston, Ed. New York: McGraw-Hill, 1975.

[7] H., H. Nagel, "Analysis techniques for image sequences," in *Proc. Int. Joint Conf. Pattern Recognition*, Kyoto, Japan, 1978.

[8] J. K. Tsotsos, "A prototype motion understanding system," Dep. Comput. Sci., Univ. Toronto, Rep. TR-93, June 1978.

[9] —, "A framework for visual motion understanding," Ph.D. dissertation, Dep. Comput. Sci., Univ. Toronto, 1980.

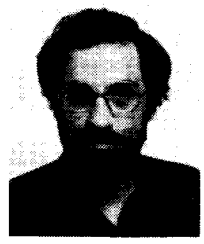
[10] S. W. Zucker, R. A. Hummel, and A. Rosenfeld, "An application of relaxation labeling to line and curve enhancement," *IEEE Trans. Comput.*, vol. C-26, Apr. 1977.

[11] S. W. Zucker, "Production systems with feedback," in *Pattern-Directed Inference Systems*, Waterman and Hayes-Roth, Eds. New York: Academic, 1978.


**John K. Tsotsos** was born in Windsor, Ont., Canada. He received the B.A.Sc. degree in engineering science 1974, and the M.Sc. and Ph.D. degrees in computer science in 1976 and 1980, respectively, all from the University of Toronto, Toronto, Ont., Canada.

He was a lecturer in the Department of Computer Science, University of Toronto, as well as a Canadian Heart Foundation Research Fellow at Toronto General Hospital for 1979–1980. He is currently a Research Consultant in Cardiology at Toronto General Hospital and Assistant Professor of Computer Science at the University of Toronto. His research interests include computer vision, knowledge-based systems, and application of artificial intelligence to biomedical image analysis and diagnosis.

Dr. Tsotsos is a member of the Association for Computing Machinery and CSCSI.



**John Mylopoulos** was born in Athens, Greece. He completed his undergraduate studies at Brown University, Providence, RI, in 1966 and his graduate studies at Princeton University, Princeton, NJ, in 1970, both in electrical engineering.

Since 1970 he has been with the Department of Computer Science, University of Toronto, Toronto, Ont., Canada. His research interests include knowledge representation and the design of knowledge-based systems.

**H. Dominic Covvey** (M'74) is a specialist in the field of medical computing. He is Director of Cardiovascular Computing at Toronto General Hospital, Assistant Professor of Computer Science at the University of Toronto, and Lecturer in both the Department of Medicine and the Department of Preventive Medicine there. His research and development activities are in the areas of ventricular function analysis, pacemaker follow-up, database management systems, and medical image analysis.

Mr. Covvey is a member of the Association for Computing Machinery and the Canadian Cardiovascular Society.

**Steven W. Zucker** (S'71–M'75) received the B.S. degree in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1969, and the M.S. and Ph.D. degrees in biomedical engineering from Drexel University, Philadelphia, PA, in 1972 and 1975, respectively.

From 1974 to 1976 he was a Research Associate at the Picture Processing Laboratory, Computer Science Center, University of Maryland, College Park. He is currently an Associate Professor in the Department of Electrical Engineering, and the Computer Vision and Graphics Laboratory, McGill University, Montreal, P.Q., Canada. His research interests include computer vision, graphics, perceptual modeling, image processing, pattern recognition, and artificial intelligence.

Dr. Zucker is a member of Sigma Xi and the Association for Computing Machinery.