Artificial Perception of Actions

Robert Thibadeau
Robotics Institute
Carnegie-Mellon University
Pittsburgh, PA 15213

Revision for *Cognitive Science*
17 February 1994

Address and phone: Robert Thibadeau, Robotics Institute, Carnegie-Mellon University, Pittsburgh, PA 15213, (412) 578-3824

Abstract

This paper has to do with the visual perception of actions that are discretely conceptualized. The intent is to develop a vision system that produces causal or intentional descriptions of actions, thus providing the conceptual underpinnings of natural langu age descriptions. The computational theory is developed in linking a "point of action definition" analysis to an analysis of how the physical events will elicit appropriate verbal descriptions. Out of this theory of direct computational linkages between physical events, points of action definition, and verbal descriptions, comes a theory of perception that provides some insight into how to go about constructing systems that can watch the world and report on what they are watching.

The aim of this research is to develop a machine that interprets visually observed events as discrete actions. Every discrete action is a momentary causal or intentional description of movement. To insure the system's veracity and usefulness, the description and thereby the "perception of action" must provide a faithful mapping to natural language. For example, the machine might verbally report that some person *is now walking toward the robot*, that a person *has left his station hurriedly*, that *a machine operator is pushing the wrong switch*, or that a person *has been walking down the corridor trying to open locked doors*, or that someone *is running* down the street. In effect, the machine must see everyday actions in a way that permits their ready communication to us.

There is hardly any active research on the problem of computing such perceptions. On the contrary, it is particularly easy to find papers that assert or strongly imply that the study of action perception must wait until the results are in on object perception and motion perception. This view is even expressed by researchers who will admit that motion perception can be studied without solving the problems of object perception.

It is my thesis that actions are perceived directly without necessarily interpreting objects or motions. I would like to understand how to account computationally for action perception without also accounting for all its ostensible precursors, like object and motion perception. Just as, in a computational sense, a motion can be perceived without first recognizing the moving object, it is possible an action can be perceived without first recognizing the underlying motions. Such a relationship should not be taken to deny indirect computation from object and motion *information*, only from the conceptualization of objects and motions. The present strategy approaches action perception intensely "bottom up" and "top down", seeking always to discover the most direct relationships that hold between the physical events and the verbal conceptions. To begin the analysis, we need to distinguish what is conventionally meant by the study of "action perception" by comparison to the studies of related perceptual phenomena such as "motion perception", "the perception of causality," and others.

Action Perception as a Study

Michotte's studies of the perception of causality (Michotte, 1963) represent some of the best known and earliest work of relevance to action perception. Michotte devoted his life to classifying physical conditions that elicit perceived causation. He discovered that perceived causality is based upon a few distinct sensations. There are only a few ways to perceive causation. By contrast, action perception is, by convention, not about distinct sensations of action, but about perceiving individual actions. The perception of causality is important in the perception of many actions, but it may not be involved in the perception of all actions.

Motion perception is often believed to encompass action perception, but the literature on motion perception deals with the perception of physical motions which are not inherently discrete. The physical motions are algebraically characterized by a cyclic or, at best, continuous, function with no naturalf beginning or ending (see Johannson, 1975; Restle, 1979; Ullman, 1979 and below). A classic problem, described in Ullman (1979) , is to understand the computation involved in merging spatially separated motions together into the perception of a single coordinated motion (e.g., seeing an arm movement as a coordinated movement of one object). Motion perception and object perception then become closely related topics. By contrast, action perception is further removed from object perception, *per se*, and is involved with discrete motions of objects. In action perception, the emphasis is as much in the conditions for the beginning and ending of motions as it is with the motions themselves.

One area of research associated with action perception is natural language understanding. A typical problem in language understanding is the understanding of natural language descriptions of actions at a "conceptual" level. There is relevant work in basic action categories (*primitives*, see Schank, 1975) and (Miller and Johnson-Laird, 1976) in which the aim is to compose complex actions out of simpler, more primitive, ones. There is also relevant natural language work on recognizing the significance of action sequences (Schank and Abelson, 1977; Schmidt, Sridharan and Goodson, 1978). The work in recognizing the significance of action sequences is of importance owing to the "hierarchical" nature of action. An action, although perceived as a single thing, is typically made up of finer actions concatenated together (e.g., body movements are a component of walking). The finer-actions can be individually perceived if the perceiver has the opportunity, and wishes, to see them.

The work from natural language understanding claims to describe a "language-free" conceptual level of representation which, by the claim, should bear on the understanding of action perception. Unfortunately, it is not that clear that "language-free" levels of representation have yet been described. Present knowledge about conceptual representation *de facto* derives from research grounded in the uses of natural language. It is true that the peripheral taxonomic structures of language, such as syntax and morphology, are distinguished from conceptual, or ideational, structures, but, from the point of view of the present research, what a linguist, or even a logician, may call ideational structures look, because of the derivation, like language structures as well. The same may be said for "conceptual representations" or "propositional representations" in AI. A major goal of the present research is to provide some independently motivated mapping or derivation between visual perception and our understanding of conceptual representation.

The study of "action perception" has been most developed by social psychologists, most prominently Heider (Heider,1958; Heider and Simmel, 1944) although I focus on Newtson (Massad, Hubbard and Newtson, 1979; Newtson, 1973; Newtson, 1976; Newtson and Engquist, 1976; Newtson, Engquist and Bois, 1977; Newtson and Rindner, 1979; Newtson, Rindner, Miller and LaCross, 1978; Wilder, 1978, 1978) . The interest in action perception stems from a struggle to understand how people attribute intention and causation to movements. Newtson’s work is of particular interest to me because it provides an empirical, objective, and observational means of coming to terms with the process of perceiving actions. In his basic paradigm, informants watch an action sequence and indicate where they perceive actions as having occurred. Specifically, they are asked to push a button every time that, in their opinion, an action occurs in the film or video-tape they watch. The researcher records the times of the button presses for later examination with regard to the activities in the film. Newtson’s "point of action definition" provides the impetus for my study of the artificial perception of actions. The premise in using a psychological basis is that people perceive actions in a fashion desired of the machine.

Points of action definition provide more than simply an idea about how others perceive actions. Most important, they provide timely clues about an ongoing process of perceiving. The alternatives, verbal reports and subjective analysis, are either not timely or just immensely difficult. A point of action definition is a simple and direct report from a person about *when* the action is seen. It turns out that the explicit point of action definition associated with verbal report and the observed physical event reveals critical relationships between the percept and the physical event. An understanding of these temporal and substantive relations is the basis for a computational understanding.

Newtson's conclusions about action perception were acquired in over ten years of research. I cannot summarize all of that research here. But one particularly relevant result is that people normally segment events into actions, even when they are not required to do so by instruction. However, people are remarkably unaware of their segmentations: A commonplace example is in film montage effects: cutting from shot to shot without confusing or making the viewer aware of the cuts. Film editors know that cuts must be made at points of action definition.

A final relevant conclusion from the research findings, is that action perception is susceptible to strong forces of expectation: viewers must be prepared to see an action in order for them to see it. The nature of this preparation is not well understood, but it is clear that a preparation or expectation system of significant conceptual impact operates in the normal perception of actions.

"Actions" in Visual Perception

This section focuses on the general problem of deriving knowledge representations of perceived actions. It begins with a conventional logical and linguistic (logico-linguistic) derivation and then develops a perceptual derivation of knowledge representations to take account of bottom up cues from the visual stream.

Top Down

There are many schemes for taxonomies of conceptual actions which attempt to simplify the representation of actions for conceptual understanding and inferencing purposes. While most all existing schemes pay little attention to how actions might be recognized perceptually, many carefully attend to the conceptual representation of actions. Schank's (1975) work with primitives in narrative prose is a familiar example and Miller and Johnson-Laird (1976) provides another. The work by Miller and Johnson-Laird is compelling because, although their technique was based on logico-linguistic analysis, their aim was to be accountable to visual perception as well as language perception.[1] To see how their logico-linguistic analysis works, and how it might work for visual perception, we can examine their analysis of some primitive actions.

The most primitive action in my system is loosely verbalized as "travel." This action was tentatively assigned the definition provided formally in (1). The notation has three parts: the proposition, **TRAVEL(x**), its presuppositions[2] (such as the presupposition that there is a place *y*), and its entailments

---

[1]Miller and Johnson-Laird (1976) provide an informal computational framework, but that computational framework will not be discussed in this paper. Their formal analysis of natural language and conceptual structure is, nevertheless, a classic contribution to our knowledge for its methodology, its depth, and its scope.

[2]"According to Strawson (1952) , statement *S* semantically presupposes statement *S'* if *S'* is a necessary condition for the truth or falsity of *S*." Page 172 in Miller and Johnson-Laird (1976), suggests a willingness to accept a pragmatic or interpersonal version of this definition which depends on belief states.

[3](as in (i)). The time indicator *t* is conventional. Miller and Johnson-Laird's use of the *statement forming operator R* (Rescher and Urquhart, 1971) is possibly novel: R_t corresponds to "this statement is recognized as true at time *t*". Another statement forming operator, Q_t, is used later and it means that "this statement is recognized as true at all moments prior to *t*".

(1) **TRAVEL**(x): Something x "travels" if there is a time *t* and a place *y* such that:

(i) Either: R_{t-i}(not**AT**(x,y)) & R_t(**AT**(x,y))

Or: R_{t-i}(**AT**(x,y)) & R_t(not**AT**(x,y))

The formulation in (1) pertains to conceptual representation, but I believe it superior to other conceptual formulations, such as Schank's PTRANS, in that it stipulates logical presuppositions and entailments for primitive actions. I will later evaluate a method of action recognition that seeks to match such semantic "necessities" against visual information. The proposition, **TRAVEL**(*x*), nevertheless, lacks the place descriptor *y*, suggesting its natural language or conceptual function. In reasoning and understanding, the particular place from which or to which the object moves is often not relevant. Miller and Johnson-Laird (1976) state through presupposition and entailment how one may elicit the perception of an action.

The statement forming operator may tell us when the conceptualization is recognized. If the conceptualization is an action conceptualization recognized at a single point in time, then we have a natural way of predicting a point of action definition. But, in fact, statement forming operators account for verb tense and aspect in Miller and Johnson-Laird (1976) and may only accidently predict points of action definition. This first version of **TRAVEL** has problems with the "point of action definition" property. The statement forming constraints specify two points of recognition, R_t and R_{t-i}. Furthermore, the entailment is *ambiguous* in point of definition: the point of definition is either the moment when the object moves from a place or the moment when the object arrives at a place.

For logico-linguistic reasons, Miller and Johnson-Laird were not satisfied with this first definition either and went on to develop their final formulation in (1'). They wanted to derive that "if x does not travel, it stays where it was." They also felt **TRAVEL** should have a durative entailment. They did not want to allow, as (1) seems to imply, that **TRAVEL** is defined when objects just disappear or appear at locations.

(1') **TRAVEL**(*x*): Something *x* "travels" from time t_0

to time t_m if, for each t_i such that

$t_0 <= t_i <= t_m$, there is a place y_i such that

R_{t_i}(**AT**(x,y_i) and:

(i) $R_{t_{i+1}}$(not**AT**(x,y_i))

---

[3]At first sight, entailments are simple conditionals: a statement *S* entails statement *S'* when *S* is a condition for *S'*. When *S* is true, *S'* cannot be false, but when *S* is false, *S'* can be true or false. However, in Miller and Johnson-Laird's usage, the negation of *S* necessitates the negation of *S'* (see page 532). By example, if it is true that "x travels" then *(1)(i)* holds, if it is true that "x does not travel" then the negation of (i) must hold as well. If either antecedent, "it is true that ..", had been false, then the consequent could have been true or false. This is a proper usage of entailment, and I will adopt it in this paper.

TRAVEL$_{(1')}$ represents a simple motion which is, conceptually, the simplest form of action. But this is still ambiguous in point of action definition because, by continuously changing the focal place, y with y_i, it now describes a primitive motion (with no fixed beginning or ending) and thereby no point of definition or, perhaps, an entirely filled sequence of them. Later I will show a way to resolve this problem and derive the primitive action from the primitive motion described in **TRAVEL**$_{(1')}$.

The sense of *primitive action* is that actions of greater complexity are built of Boolean predications which directly or derivatively employ the primitive predicate. The predicate **MOVE** from the English transitive verb "move" is the primitive **TRAVEL** made complex by a perception of agentive **CAUSE**. The verb "move" also has an intransitive usage, as in "he moves", which is not addressed here. Rather, we have:

(2) **MOVE**((x),y): Something x "moves" something y if:

(i) **TRAVEL**(y)
(ii) **DO**(x,**S**)
(iii) **CAUSE**(**S**,(i))

The **DO** is, in Miller and Johnson-Laird (1976) and the views of many others, a placeholder for other movements, This placeholding is not without presumed consequence. It is a further manifestation of action "hierarchies," in this case a form of one action (the **TRAVEL**) subsuming whatever action(s) composed the **DO**.

The relationship between the **DO** and the **TRAVEL** is potentially more complex than a **CAUSE** (Michotte, 1963 can be consulted for an appropriate perceptual analysis of causation). If the action is perceived as intentional, as this sort usually is, then the subsumptive relation is an **IN-ORDER-TO** relation: the actor did something in order to move the object. The **IN-ORDER-TO** relation has been worked out by analysis similar to Miller and Johnson-Laird (1976) in Schmidt and Sridharan and Goodson (1978).

The added entailments in (2) also add complexity to the question of resolving points of action definition. Causation, for example, is perceived at a moment in time and could predict a point of action definition. However, other complex actions establish points of action definition in explicit ways and without causation. Consider the presuppositions and entailments for "reach" and "depart" in (3) and (4). The actions are momentary, at t, and each action presupposes a past, durative, activity as indicated by the statement forming operator Q_t.

(3) **REACH**(x,w): Something x "reaches" some place
w if there is a moment t such that
Q_t((**TOWARD**(**TRAVEL**))(x,w)) and:
(i) R_t(**AT**(x,w))

(4) **DEPART**(x,w): Something x "departs" some place
w if there is a moment t such that
Q_t(not**TRAVEL**(x)) and:
(i) R_t(**TRAVEL**(x))

With these formulations, the referential ambiguity in **TRAVEL** is reduced by additional statement forming and presuppositional constraint. The point of action definition is suggested as the moment, t, at which the actor reaches a location or begins to travel.

The illustrations with the "move," "depart," and "arrive" demonstrate a taxonomic system for action we would like to have available to processes of action perception. However, there is a failure in the logico-linguistic analysis to support such a scheme for perception. We must look elsewhere for that analysis.

Bottom Up

Another method for approaching action definition carries us further from natural language considerations and closer to deriving conceptualizations from physical events. This previously unpublished definition was formulated by a colleague, John Rotondo, working with Newtson to develop mathematical (not computational) models. To set the stage for this method we make a simplifying assumption that it is possible to describe the effective input into the perceptual system as a sequence of states, **S**; where a state, s_t, is an interpreted image[4] at a moment, t, in time.

The generative method recognizes two classes of change in a succession of states, **S** at s_t: a simple state difference or first-order change, c_t^1, and a change in a change or second-order (derivative) change, c_t^2. The number of moments contributing to these changes is not specified: successive state differences can compute a constant curving movement, a uniform acceleration, or constant cycle, as well as straight movements and constant velocities. A change in a change can just as well be a change in direction as a change from movement to no movement.

A generative description of the domain of possible descriptions of an event discretely continuous over time is obtained by evaluating all change descriptors for all moments in time. This generative description, it is claimed, *generates the superset of all actions in an activity*, as well as all motions. It takes little imagination to recognize that the number of generated actions and motions rapidly grows very large. Even with dramatic simplification, such as differencing between just two successive states, real world imagery produces enormous numbers of change descriptions over short periods of time. We can, however, now view action perception as problem solving and formulate action schema in terms of a search method through a space of descriptions.

Three reductions are used to limit the search:

- Points of action definition are drawn from the collection of second-order changes, $\mathbf{C}^2$. This is called the *action postulate*.
- A *systemization of schemas* which recognize actions is provided. This is similar in spirit and form to Miller and Johnson-Laird (1976) .
- A system for dynamically elaborating schema and making selected schema available to perceptual processes is provided called the *expectation mechanism*.

The action postulate. The action postulate is advantageous when it is important to monitor changes, such as a continuous movement toward a place, over some period of time, just as one monitors states, such as an object at rest. In describing the perceptual activity, as opposed to the conceptual activity, monitoring environmental invariances becomes important (Gibson, 1966).

---

[4]Two, two and a half, or three dimensional segmented, and perhaps conceptually labeled, representation of the visually encompassed world for a moment in time.

The *action postulate* was derived empirically.[5] used the term "feature of change" rather than "second-order change" in his description of the postulate.). Newtson's studies (some of which are not published) have suggested that the action postulate, while correct in one sense, is only approximately precise. By example: In a film of a man hammering a nail into a block of wood, people routinely report the action slightly after the second-order change has occurred. People differ in selectivity: some will report actions only when the man hits the nail while others will see the man raising his hand as an action as well. Nevertheless, they are consistent in reporting the action several instants *after* the hammer stops on the nail or *after* the hand is at its full altitude and just begins to fall. My own observations have confirmed such a delay. The only work with the delay has used film speed variations to confirm that the delay is not simply reaction time delay: As film speed is increased, the delay is shortened, and, as it slows, the delay lengthens. Unfortunately no systematic investigation of this delay for different classes of action has been undertaken to date.

In the next section, we develop a system to represent the action postulate explicitly and account in general terms, for the observed imprecision in the postulate. The system derives particular action schema from a parent action schema in a fashion analogous to the systemization in Miller and Johnson-Laird.

Action schemata for perception. The point in time where an action occurs provides concrete references for the undertaking in artificial perception, but as a practical computational matter, it is good to know how the machine is enabled to see actions defined at particular moments in time, from bottom-up cues. The cues I currently allow are the state descriptions, **S**, and first-order change descriptions, $\mathbf{C}^1$, available to perception.[6] The major part of the computational study was in how to define action schemata which would map these bottom up cues to action conceptionalizations.

I initiated the computational study without any presumption of a single primitive schema for actions. In fact, I set out to develop any number of elaborate finite state machines consistent with Newtson's observations that could be employed to recognize different actions. These state machines were found to be largely compositions of a simpler machine which I now think of as a primitive or parent action schema. This parent action schema and its systemization is the topic of this section.

The parent schema is defined as an automaton, thereby defining the schema in procedural terms, in the same spirit as Piaget's, (1963) formulations . The systemization asserts that *every action perceived is a manifestation of a modified instance of the parent action schema. In that manifestation, this instance has been evaluated against the sensory stream and has thereby successfully completed the instantiation of its associated action conceptualization*.

The parent action schema has five parts which includes the declarative conceptualization which the automaton functions to instantiate and assert, two special conditions it monitors in the sensory stream, a procedural method for robustness against noise, and a procedural subsumption method for forming compositions of schemas which would signify a single conceptualization. These components, their modification and evaluation, are described in detail below:

---

[5]Newtson (1976

[6]For this paper, I set aside discussion of recursively allowing second-order change descriptions to influence the computation of new second-order change descriptions, as in "The man repeatedly went to get coffee".

- *A Conceptualization*,CON, to stand for the action at the moment it is conceptualized.

- *A Criterial Component*,COLLECT, which is a Boolean composition of simple state and first-order change descriptions to be detected in the sensory stream. This component is not the only referential component criterial to an action, but it is the only referential component criterial to the point of definition of the action. All actions will be defined at the point in time the criterial component goes from true to false by reference. For example, a person breaking a glass might include monitoring the fall of the glass toward the floor. When the glass ceases falling, or perhaps moving (within a limited pattern), the action is defined.

- *A Non-Criterial Component*,ACHIEVE, which is a Boolean composition of state and first-order change descriptions to be detected in the sensory stream. This non-criterial component is consulted before and at the moment of action definition to determine whether the action itself is true or false (whether the glass broke or bounced). While this component is referentially criterial (it determines the truth of a proposition on referential or deictic grounds), it is not criterial in identifying the point of definition of the action.

- *A Perceptual Error Correction Component* which determines the tolerance in evaluating the change in truth value of the Criterial Component. This component represents a system of processes which modulate perceptual decisions: For example, it would guarantee that a brief stop in movement (owing to any error or weakness of man or machine) would be routinely ignored and not signal a false point of action definition.[7]

- *A Motion Linking Component*, NEXT, which permits chaining action conceptualizations (viz., CON pointers) for "macro" definitions. This is an optional component. The presence of a NEXT predication can suspend a true-false decision through the ACHIEVE Component, and thereby it can suspend a point of action definition. An example of the use of the NEXT component can be found in one of two ways of viewing a man hammering a nail (see Newtson, Rindner, Miller and LaCross, 1978): (a) simple, one schema, actions such as the hammer hitting the nail or the hand rising, and (b) a sequence of the hand rising then the hammer coming down and hitting the nail. The sequenced perception may be of use in stages of becoming skilled

[7]The specific error correction system generally depends on the details of the implementation. In the implementation described later, the system associated two parameters with each action schema, (a) a minimum duration threshold for the COLLECT proposition to turn true (after verification begins) and (b) a maximum duration threshold over which the COLLECT proposition was not true before deciding the COLLECT proposition was in fact not true.

at perceiving. If the hand rising has no other effect than to allow it to fall, the action schema may be shortened to contain only the last part of the hammer hitting the nail. A further reason for the NEXT is that the processes which adjust what actions are "seen" will no longer posit "the hand rising" as a relevant thing to monitor once the NEXT relationship is recognized between the rising and the hitting.

Instances of action schema can be derived for Miller and Johnson-Laird's (1976) examples. I will use a one state proposition, **At**(x,y), which signals that an object, x, is at a place y, and one first-order change proposition, **Move**(x) which signals that an object has changed its position in space. To recognize a $\mathbf{TRAVEL}_{(1)}$ requires at least two primitive instantiations of the parent schema to conform with the ambiguity in its entailment. The instantiations are given in (5) and (6). The new representation fills designated slots in the schema described above.

(5) CON: $R_t(\mathbf{TRAVEL}_{\mathbf{(1)}}(x))$
COLLECT: **At**(x,y)
ACHIEVE: not**At**(x,y)

(6) CON: $R_t(\mathbf{TRAVEL}_{\mathbf{(1)}}(x))$
COLLECT: not**At**(x,y)
ACHIEVE: **At**(x,y)

The action postulate says that the point in time, t, that the COLLECT for (5) or (6) goes from true to false, there is a determination of the truth or falsity of the conceptualization, CON. In these cases, the conceptualizations, if ever recognized, will be true, since the COLLECT is the inverse of the ACHIEVE. A more serious problem with this formulation is that it violates the action postulate since it only detects a first-order change (change between two states), not a second-order change (a change in a change). So, like Miller and Johnson-Laird (1967) we are led to reject this formulation as unsound. $\mathbf{TRAVEL}_{(1)}$ conflicts with the action postulate.

Two different schemas are required to recognize the durative form of $\mathbf{TRAVEL}_{(1')}$. It is not possible to report the ongoing change at an arbitrary moment: we must select the beginning of the travel as in (7) or the end of the travel, as in (8).

(7) CON: $R_t(\mathbf{TRAVEL}_{\mathbf{(1')}}(x))$
COLLECT: not**Move**(x)
ACHIEVE: **Move**(x)

(8) CON: $R_t(\mathbf{TRAVEL}_{\mathbf{(1')}}(x))$
COLLECT: **Move**(x)
ACHIEVE: **Move**(x)

These schemas collect the absence of a change of position or the change of position. When either of these conditions change (go from true to false), the system reports the object, x, has traveled. The apparently unambiguous $\mathbf{TRAVEL}_{(1')}$ has two forms, like $\mathbf{TRAVEL}_{(1)}$, but because of the imposition of a natural point of definition on the action of traveling. These, nevertheless, represent primitive conceptual actions by perceptual derivation. $\mathbf{TRAVEL}_{(7)}$ is virtually the same in semantic form to **DEPART** in (4). However $\mathbf{TRAVEL}_{(8)}$ corresponds to a past tense version of primitive movement $\mathbf{TRAVEL}_{(1')}$.

The current system provides for the delay between the physical moment, t-i, when the COLLECT proposition goes from true to false, and the response moment *t*, when the point of definition occurs, in the simplest fashion possible. Because of the effect of noise, the perceptual error component allows some absorption in deciding whether the COLLECT proposition has gone from True to False. I assume that the same absorptive process that keeps the COLLECT proposition True in noise also accounts for the observed delay (perhaps as much as several seconds) in reporting that an action has occurred. Since the COLLECT monitoring rate is defined by reference to the environmental frame rate, not in absolute time, this explanation is also in agreement with the preliminary experimental findings when film projection rates are speeded or slowed.

Defining action schemas for recognition is clearly different from defining them for conceptual analysis, but the taxonomic principles of schema organization are potentially similar. Instantiations and compositions of the instantiations[8] of the parent schema provide the basis for taxonomic categorization of action schema. Since conceptualizations are a product of schema recognition, the taxonomies can naturally reflect the taxonomies of conceptualizations. But this sort of taxonomic constraint is definitional and not very interesting.

It would be interesting if we could also carry down the conceptual constraints (such as the "toward" modification on **TRAVEL** which contributes to **REACH** in (3)) to the COLLECT and ACHIEVE levels of appropriate action schemas. Formally, COLLECT carries presuppositions and ACHIEVE carries entailments: a True COLLECT proposition is required even to entertain the truth or falsity of the CON. But given that COLLECT is True, a True or False ACHIEVE proposition determines whether the CON is True or False. This is an important connection between the linguistic and perceptual analysis.

The match also depends upon the first-order change and state descriptors that result from bottom up processing. As a research tactic, these descriptors can be adjusted to guarantee compatibility. The most dominant factors in carrying out conceptual constraints are, then, the expectation mechanism and the presence of the required information in the physical events. These are the next topics.

Expectation Mechanism. The model of action perception requires the generation of pools of action schemas to act as independent finite state automata that monitor the state and first-order change information on a moment by moment basis. Action schemas monitor derivative information, not the physical events directly, in as much as their function is to respond to second-order changes.

The relationship between action schemas and state and first-order change information is fundamentally a matching relationship where current action schemas are repeatedly matched until the COLLECT proposition in any one goes from true to false and an action is defined. The *expectation mechanism* has the role of identifying the current subset of action schema which are, for one reason or another, worthy candidates for such matching. The expectation mechanism may both *generate* and *select* action schemas.

There are reasons to limit the set of current action schemas to a subset of action schemata.

- To do a physical implementation which operates in something close to real time, it is of course necessary to limit the set size.

---

[8]Using the NEXT mechanism.

- Action schemas are all instantiations and compositions of instantiations of a most general action schema. There is no requirement for complete instantiations. For example, there might be a general "open a door" schema that matches anyone opening a door. If two schemas are available which differ only in specificity, then only the least specific schema requires matching. I know of one empirical observation that supports the availability of incompletely instantiated action schemata: If something unexpected occurs in an activity, the temporal density of points of action definition increases abruptly (Newtson, 1973; Newtson and Rindner, 1979). This effect has been called a *primacy effect*; it is more like an reorientation effect. It suggests that people initially use more abstract action schema to pick up on the "micro-actions" in order to make initial sense of what is going on.

- There are logical constraints on schema relevance: an "open a door" schema need only be monitored when a door is in the picture. There needs to be a mechanism which obviates the need for a 'close the door' schema to be monitored when the door is already closed.

- In addition to the relevance of partial matching and logical constraints, perceptual confusions are a functional reason for selectivity. Preliminary empirical work suggests this functional basis may be a most significant one. Over the long term, a variety of context mechanisms are important to reduce the complexity of action schema. The fact that a COLLECT proposition represents any arbitrarily complex Boolean predication of state and first-order change descriptors does not assure that any arbitrarily complex Boolean predication of such descriptors is in fact sufficient to distinguish one action from another. Perception is selective in profound ways: the same activity and the same physical manifestation may have different significance on different occasions. It is well recognized in Epistemology that one can take virtually any action one can conceive of and make up an example: "closing the door" may alternatively be seen as "pushing the door as far as it will go" (with exactly the same schema except for the CON designation). The difference may be in some higher intention, in the example, containing something versus getting the door out of the way. The NEXT component was designed specifically for this reason. In the hammering example, raising the hammer then (NEXT) bringing it down seen as a single action by a perceptual sequence may, with further accommodation, be seen as hitting the table with the hammer. Also recall the **DO** in linguistic semantics: Raising the hammer and lowering it is what the actor **did in order to** hit the table. This problem has been extensively addressed from an AI perspective by Schmidt and Sridharan (Schmidt, Sridharan and Goodson, 1978; Sridharan and Schmidt, 1978). I have, in effect, adopted their resolution which posits the use of an expectation mechanism.

A reasonable framework for talking about an expectation mechanism is in terms of a generalized rule system, or production system (Thibadeau and Just, 1982). The function of the rules in the production system is to adjust the expectation set. The only adjustment which can be made is the addition or subtraction of action schemas to and from the expectation set.

Within this framework, I propose a new postulate, the *expectation postulate*: *the expectation set is adjusted only at points of action definition*. The rules in the production system add or delete action schemas from the expectation set in response to actions. One of the reasons for identifying a production system as a useful preliminary model for an expectation generator is that it is structured to respond quickly to new demands (Erman and Lesser, 1978; Tanimoto, 1982). The information available to the rules in the production system can be considerably richer than simply the most current action recognized and the set of all action schemas. Furthermore, the rules can posit intermediate results and can apply with intermediate effects other than simply the addition and subtraction of action schemas for the action recognition matcher.

A potential practical problem with the expectation postulate is that artificial perception could be effectively blind to unexpected actions at least for a significant period of time. To reduce blindness, a proper role of the action recognition matcher would be to carry an implicit action schema that can signal significant activity not being accounted for by any action schema. This could be given by an action such as $\textbf{TRAVEL}_{(7)}$(x) signaled at the beginning of a movement with a free variable for x.

From one perspective, the action $\textbf{TRAVEL}_{(7)}$(x), once recognized, is treated like any other recognized action: the rule system provides additions and subtractions of action schema in response to the action. It is likely that having the "wired in" accounting mechanism alone is more desirable than a "wired in" $\textbf{TRAVEL}_{(7)}$(x) schema. An explicit loading of situationally sensitive $\textbf{TRAVEL}_{(7)}$-type schema is preferable. Such schemas do not respond unless the second-order change they detect is not accounted for in any other action schema (they always compute off residual state and first-order change descriptions).

A Complete Model

Having a parent action schema and an expectation mechanism does not tell us the content of schema instantiations and how to compose them. Logico-linguistic methods seem appropriate for filling out conceptual representations. But a different method is called for in action perception since the problem is one of defining the perceptual engine, not manipulating the content it generates. I believe that empirically oriented "protocol studies" are the most useful: informants provide not only their English descriptions of the actions, but also an indication of points of action definition. Action schemas can then be tailored according to human demands in a complete computational model which goes from raw visual imagery (e.g., Cartesian data or pixel data) to conceptual description.

I had an opportunity to do one such analysis which included a protocol analysis and experiment, and the results of that study are promising. An old animated film was selected for the study. The film was constructed on a "flatland" theme by Heider and Simmel (1944) . The film has been extensively discussed, and peoples' perceptions of it extensively researched, at least since 1944 (Greenburg and Strickland, 1973; Heider and Simmel, 1944; Massad, Hubbard, and Newtson, 1979; Shor, 1957.) The story-line of the film is not trivial, although the characters are two triangles, a circle, and a box with a hinged door on it. Appendix I has a picture). Let me reconstruct the story in English terms (since it is not possible to faithfully render a movie in a page of print). In this reconstruction, I have tried to be faithful, in flavor, to how one might first view the film (for example, there is a vivid impression of the gender of the actors, which is retained in the text):

> A large triangle is just inside the box at the open door. He closes the door and goes to the bottom of the box.
>
> A small triangle and a circle enter the screen together. The small circle stops out front of the box while the small triangle heads around until he is at the door.
>
> The large triangle then goes to the door, opens it, and, almost immediately, he viciously attacks the small triangle. The small triangle is repeatedly hit against the wall of the box and finally goes under the box in apparent submission.
>
> The large triangle now goes after the small circle who has hidden in the box and closed the door. The large triangle gets into the box and closes the door behind him. He now tries to corner the small circle but she never gets hit or pinned down. Rather, the small triangle opens the door and the small circle escapes.
>
> The small triangle and small circle appear to kiss while the large triangle has difficultly trying to open the door which is apparently locked. The large triangle, on getting out of the box, chases the twosome around the box but they escape off screen.
>
> At this point, the large triangle moves over to the door, opens and closes it, and then hits it, breaking the door then the box into tiny pieces.

I asked three students to watch the film several times each. The result of the analysis of the protocols and the physical events in the film is a total of 64 action schemas with no variables for objects. The physical action schemas were applied with simple (logical and thematic) expectation adjustments to the film to yield a total of 149 instantiation points of action definition over the film.

Having the characters depicted as simple geometric forms in 2D space considerably simplifies many image analysis computations and allows us to study action perception rather directly. Note that although the movie shows only geometric shapes, the perception of sentient, thinking, characters is as vivid as in any cartoon animation. As a narrative, this cartoon is not at all simple. Actions by the actors are not perceived as simple movements on the screen, but as implication-rich intentional actions. There are examples of cooperation, vicious attack, submission, planning, surprise, and even "kissing". The story structure is a standard narrative form: the paragraph structure in the description correspond to the elements: (1) Setting , (2) Motivation, (3) Primary Engagement, (4) Primary Engagement, (5) Secondary Engagement, (6) Value Resolution.

In my computer analysis, I used the cartesian coordinates of all the line segments and arcs in every other frame of the 1690 frame film as the raw data about the physical events. A uniform computation of state descriptions and first-order change descriptions over the entire film was made from the Cartesian description (see Appendix I). The state and first-order change descriptions were generated by a FORTRAN program. A LISP program was used to analyse the theory of action perception. The protocol analysis involved comparing the human responses against the computed state and first-order change description, formulating action schemata to render those responses as faithfully as possible, and running the LISP program to confirm the results were reasonable. Of the nine months on this project, this work represented eight months.

The state and first-order change descriptions from the program for frames 538-542 are shown below. Frame 540 corresponds to the moment before the informants said that the large triangle intentionally hit the small triangle. Since we know a **HIT** was seen, the state and first-order change descriptors which first appear relevant have been printed in bold face. The rendition reflects the fact that the description at this level will omit a predication if it is referentially false: this was possible because we used the finite and deterministic referential language described in Appendix I.

(9)Computer generated State and First-order change descriptors for three successive frames.

Frame 538: At(<all actors>,outside-box)
At(small-triangle,wall)
At(small-circle,door)
Door(Open)
Near(small-triangle,large-triangle)
Line-of-sight(<mutual among all actors>)
Move(large-triangle,normal-speed)
Move(small-triangle,slow-speed)
MovePast(<large and small-triangle>,small-circle)
MoveToward(<large and small-triangle>,wall)
MoveFrom(<large and small-triangle>,left-border-of-screen)
MoveToward(large-triangle,small-triangle)
MoveFrom(small-triangle,large-triangle)

Frame 540: At(<all actors>,outside-box)
At(small-triangle,wall)
At(small-circle,door)
Door(Open)
At(small-triangle,large-triangle)
Line-of-sight(<large-triangle and small-circle>,small-triangle)
Move(large-triangle,normal-speed)
Move(small-triangle,normal-speed)
MovePast(<large and small-triangle>,small-circle)
MoveToward(<large and small-triangle>,wall)
MoveFrom(<large and small-triangle>,left-border-of-screen)
MoveToward(large-triangle,small-triangle)
MoveFrom(small-triangle,large-triangle)
Move(small-circle,slow-speed)
MoveToward(small-circle,<entrance-to-box and large-triangle>)
MoveFrom(small-circle,top-border-of-screen)

Frame 542: At(<all actors>,outside-box)
At(small-triangle,wall)
Touch(small-triangle,wall)
At(small-circle,door)
Door(open)
At(small-triangle,large-triangle)
Line-of-sight(large-triangle,<other two actors>)
<there is no movement>

It must be emphasized that the perception that the large triangle intentionally struck the small triangle is very strong despite the fact that the two triangles do not touch each other. Certainly *contact* is presupposed (and entailed) in hitting, yet the perceptual criteria appear less demanding. Note that actions are viewed from above, so physical interactions are not hidden from the viewer.

Other examples of hitting also suggested the need for weak, underspecified, definitions. The schema which correctly detects the 12 cases of hitting is shown in (10). Roughly speaking, the action is defined in a brief moment of apparent contact in which the aggressor is moving toward the aggressee.

(10) CON: **HIT**(aggressor,aggressee)
COLLECT: MoveToward(aggressor,aggressee) AND
At(aggressor,aggressee) AND
(MoveFaster(aggressor,aggressee) OR
notMoveToward(aggressee,aggressor))
ACHIEVE: At(aggressor,aggressee)

On another point, the Touch or contact predicate in Frame 542 was notated independently of the Cartesian description by the secretary who transcribed the film off the film editor. It was inconsistent of her to describe the small triangle touching the wall in frame 542 but not frame 540. This is simply a source of error which occurs since the various state and first-order change predicates can derive from different processes in the input stream. A recognition system that cannot routinely compensate against such minor errors is not interesting. Error in description certainly contributes to the lack of specificity in schema, but I confirmed that the film also lacked the necessary information.

As with the hitting schema, a number of other actions reported by the respondents were carried out more than once. A few of the other more commonly recognized action schema from tdhe computer analysis (discounting for actor instantiation as in (10)) are provided in (11-13). Again, the recognition criteria are weaker than might be supposed by an analysis of entailments and presuppositions:

(11) CON: **OPEN**(person,door)
COLLECT: Door(Opening) AND
At(person,door)
ACHIEVE: Door(Open)

(12) CON: **SHAKEorROLL**(person)[9] has reference:
COLLECT: Rotate(person) AND
notMove(person,normal) AND
notMove(person,fast)
ACHIEVE: Rotate(person)

(13) CON: **WENT**(person,door,(from inside box))
COLLECT: At(person,inside-box) AND
MoveToward(person,door) AND
notTouch(person,door)
ACHIEVE: At(person,door)

The actions in (11-13) are all distinct from the **HIT** in that they are durative. This provides for a small adjustment in the perceptual error component and that in turn agreed with the observed delays in the points of action definition. Furthermore, the point of definition for **HIT** is at the moment of causation, while the point of definition for the other actions is at a moment when a relevant movement ceases.

---

[9]The respondents used the language "shake or roll". So will I.

The translatory movement in (13) could have been generalized by eliminating the "from inside box" criterial feature, but there was a tradeoff between specificity and sheer numbers of points of action definition. A very general translatory schema provides many hundreds of points of action definition, even in this film. Nevertheless, what was compelling in the analyses was the fact that the actions predicted by the action schema matched reported points of action definition with precision good to approximately three Frames (approximately 1/8 second). The study leaves uncertain, however, whether it will be possible to describe action schemata in terms of presuppositions and entailments from corresponding English expressions. It does suggest that a unification of action schema taxonomies and action conceptualization taxonomies is hard to achieve.

Experiment: Plan Recognition or Local Intentionality?

Although the machine implementation included an explicit expectation mechanism, the protocols could not reveal much of its character. An experiment was undertaken to sort the 149 points of action definition in ways that might give us an indication of the nature of the expectation mechanism.

Action perception for intentional actions, the predominate class of action in the film, is often discussed as a function of plan recognition. Plan recognition has to do with specifying how to recognize the plans, beliefs, and goals of actors on the basis of their actions. The seminal AI work in this domain was done by Schmidt, Sridharan and Goodson (1972), Sridharan and Schmidt (1978), Schank and Abelson (1977), Bruce and Newman (1979), Solway and Riseman (1977), and Wilensky (1978, 1978). ,(Schmidt, Sridharan and Goodson, 1972; Sridharan and Schmidt, 1978, Schank and Abelson (1977) , Bruce and Newman (1979), Soloway and Riseman (1977) and Wilensky (1978,1978). The general supposition made by these researchers is that action perception is often explicable in terms of plan recognition. In other words, the actions attributed to an actor have to do with the beliefs one holds about the actors' intentions.

Schmidt and Sridharan noted the selective nature of action perception and developed elaborate schemes for adjusting expectations against knowledge developed in plan recognition. However, there has been no empirical assessment of such an expectation mechanism for action perception, *per se*. I assume that the case is made that plan recognition considerations are relevant to the expectation mechanism (further supposing an expectation mechanism is worth considering in the first place), and that there is interest in weighting the relative significance of different factors.

The 149 points of action definition generated by the protocol analysis represent a small percentage of all the actions that could have been specified, but it is fair to assume that the collection represents a sample biased toward the most likely actions consistent with the present theory. I sought to avoid further bias by testing two alternative hypotheses. These hypotheses adjust the probabilities of action perception differently for the qualitatively different actions. The points of action definition were those collected by Massad, Hubbard, and Newtson (1979) by 55 students watching the film for the first time. The reader should consult that article for the exact methods used in data collection. Newtson's studies have suggested that the only critical aspect of the method is that people are told to "push the button whenever, in your opinion, an action occurred." The experiment concerns the probability distribution of button presses over time by predicting probabilities of the actions reported by the machine in the protocol study. This was an experiment in the best sense: despite almost nine months of opportunity, during the machine implementation and protocol study, I did not look at the true probabilities of points of action definition until after the assignments of machine interpreted actions into hypotetical probability categories was made.

The alternative hypotheses both assume that actions are seen in terms of an actor's goals, but they weight probabilities of seeing an action by different kinds of goal significance. It is supposed that the probability of an action being reported is either

- due to the natural hierarchies of planning, viz., subgoal-goal structures for an actor's plans (*the plan hypothesis*), or
- the subjective certainty that the goal was the actor's goal, viz., the observer believes the actor has the goal (*the belief hypothesis*).

Neither of these hypotheses is highly quantifiable given present knowledge, but they do readily permit an assignment of the 149 actions into two or three probability classes.

The Plan Hypothesis. The higher-level plan structure of the film seems obvious and provides for competing goals of the actors. Shor(1957) confirmed this was true for 95% of the people who viewed this film.

1. The large triangle wants to beat up the small triangle and rape or beat up the small circle.
2. The small triangle and small circle want to prevent that.

Just to make sure, Newtson told the 55 people before viewing the film that it was about a bully and two innocent passerby. In any event, the major goals are evaluated through a *composition* of subgoal successes and failures in the film. The large triangle succeeded on occasion in beating up the small triangle and in at least successfully threatening the small circle. The small triangle and circle succeeded on occasion (and finally) in preventing this. A success for them was always a failure for the large triangle (except for a short episode of success in cooperative planning which dealt with the evident affection the small triangle and circle had between themselves). Specifically, the plan hypothesis generates three classes of action (in decreasing order of expected probability):

1. **LT-succeed**. The large triangle succeeded in achieving a subgoal which directly supports the supposition that his major goal is achieved. (For example, the large triangle's continued hitting of the small triangle conveys a sense of the large triangle succeeding with each hit.)
2. **ST&SC-succeed**. The small triangle and small circle achieve a subgoal directly supporting the supposition that their major is achieved. (For example, the release of the small circle from the box by the small triangle conveys a sense of the small guys winning.)
3. **Irrelevant**. The action does not directly support either supposition. Several of the door openings and closings are of this sort: they convey no sense of success or failure.

Lest the reader be misled by the presentation: the above breakdown would also occur in a subgoal-goal structuring of the actors' plans using formally defined **IN-ORDER-TO** links (see above and Schmidt,1978): Relevant actions are put in immediate **IN-ORDER-TO** relation to the major goals, while irrelevant actions are either not related at all or are at least one action removed from the action **IN-ORDER-TO** achieve the major goal. Since plan recognition structure is a dynamically changing knowledge representation, multiple final achievements are possible if we look at the perceptions on a moment by moment basis.

The Belief Hypothesis. The belief hypothesis has to do with the perceived clarity of the action: "How clear is it that the actor *intended* to do that particular action?". A number of studies have confirmed the importance of the observer forming such belief structures for plan recognition and story understanding. In contrast to the plan hypothesis which can be appreciated from an analysis of verbal expressions as well as physical events, it is difficult (and I think, impossible) to appreciate the belief hypothesis from an analysis of verbal expressions, since natural English is so laden with supposition. It is easier to appreciate this hypothesis if one visualizes assigning evaluations to physical events -- which is, of course, what this research is all about. Again I identified three categories of action in decreasing order of probability:

1. **Unambiguous**. The actor clearly wanted to ACHIEVE the state or first-order change specified in the CONceptualization. Hitting actions and door manipulations (door closing, opening) are usually instances of such actions.

2. **Ambiguous**. Many actions, such as most simple traveling actions, convey ambiguity or no clear sense that the actor wanted to ACHIEVE a particular state (did the actor want to leave the place or go to another?). Such actions are sometimes seen in retrospect: we see some movement, but its significance (viz., the *action*) does not become apparent until later.

3. **Unrelated.** Some few actions were unrelated to any actor's goals: For example, leaving the screen cannot be a goal of an actor, since the actor does not know where the screen is.

As with the previous hypothesis, this one is directly related to a rigorous formulation (see Bruce and Newman [2]). Also like the previous hypothesis, we found that this classification of the actions against the film is a natural one.

Results. For each hypothesis, a multivariate regression analysis was used to predict the probability of a point of definition for the one second intervals over the 71 second film. The use of this analysis is quite straightforward. As in any regression analysis, the values of the dependent variable are predicted by an equation which weights the values of one or more independent variables. The regression analysis will derive an optimal measure of correlation, or best-fit, between the values of the independent variables and the values of dependent variable. In our study two regression analyses, are compared against each other to see which best fit accounts for when people see actions. The analyses have the same dependent variable but different independent variables.

The dependent variable for both analyses is the observed probability of a point of an action definition. This takes on a value for each one second interval over the film. Newtson obtained the 71 dependent values on that variable by summing the button presses by all 55 respondents within each interval. These sums are shown in Figure 4-1. This distribution is typical of distributions obtained in other Newtson's studies. It is obvious that points of action definition are well agreed on by the 55 respondents and that the distribution of points of action definition is highly variable and therefore interesting. A single person only rarely indicates two points of action definition in an interval, so this dependent variable very closely approximates the probability that a person will report an action in a one second interval. It is the function of the regression analyses to estimate how well the hypotheses will predict when and with what probability a person will report seeing an action.

The independent variables for each of the two regressions was a count of the number of points of action definition provided by the machine reports for each corresponding one second interval. The machine can provide many such points of definition in a single second, whereas a person is physically limited (by reaction times and perhaps decision processes). Therefore the regression was done as a quadratic approximation to correct for asymptotc, non-linear, response characteristics. The quadratic polynomial approach was planned before the data was seen. By using a quadratic model (which squares the independent variable) we *allow* that this may mean the person presses only once despite rapid perceptions, although the model will still reveal the linear component since it contains the nonsquared version of the independent variable as well.

The Plan Hypothesis and the Belief Hypothesis represent distinct models and are evaluated in separate regressions. Each hypothesis classifies actions reported by the computer program into three mutually exclusive categories for each one second interval. Since each category has a squared and a nonsquared version, this provides six independent (linearly weighted) variables to predict the values of the dependent variables.

The regressions were done in a stepwise fashion to give an indication of which variables are independently predictive of observed points of action definition. The order of entry of the variables with cumulative variance accounted for in the dependent variable are:

- *Plan Hypothesis*:

    1. LT-Succeed (18%)
    2. ST&SC-Succeed (21%)
    3. ST&SC-Succeed^2 (24%)
    4. LT-Succeed^2 (25%)
    5. Irrelevant^2 (26%)
    6. Irrelevant (26%)

- *Belief Hypothesis*:

    1. Unambiguous (33%)
    2. Unambiguous^2 (53%)
    3. Ambiguous (54%)
    4. Ambiguous^2 (54%)

Figure 1: Frequency Distribution for Points of Action Definition over the Heider and Simmel Film.

5. Unrelated (54%)[10]

These results are clear. Both hypotheses are successful in predicting the distribution of points of action definition shown in Figure 1, but the belief hypothesis accounts for about 100% more of the variation than the plan hypothesis. In fact, we can collapse the success in prediction to a single factor. The unambiguous action factor of the belief hypothesis alone accounts for most variance (53%).

An examination of where the plan recognition hypothesis failed and the belief hypothesis succeeded suggested that the failure occurred primarily because physical-object-related actions like opening and closing a door (but not moving from one place to another)despite their merely instrumental roles in the actors plans had high probabilities of eliciting a point of definition. This characteristic alone seems to have made the difference in the two situations.

These results have an impact on the design of an expectation-based system for action perception. The success of the belief hypothesis suggests that expectation rules should work with knowledge of the actors and objects in the situation. This knowledge should permit a simple-minded assessment of intentional actions. It may be possible, in many applications, to formulate simplified rule systems which rely on heuristics and logical constraints to adjust the expectation set. The heuristics would focus less on the plans and ultimate goals of the actors than on locally relevant actions -- what goals an actor is likely to have in a physical situation. Thus, in a room with a door, the system may posit door closing and opening routinely regardless of hypotheses about actors' plans. Perhaps the visual system should be set up to detect obviously intentional actions without regard to current hypotheses about what an actor is trying to accomplish. This is analogous to a lexical lookup as that process is employed Schank's (1975) request-based parsing. In certain situations, it may be possible to use the objects in view to index action schemas. The observations suggest that people are more simple minded in the act of perceiving than they might be in reflecting on what they have perceived. Perhaps plan recognition is more a phenomenon of *explaining* actions, as Wilensky (1978) suggests, than for preparing for them. Such results could be taken as a caveat not to try to make artificial systems for action perception too smart, since the human existence proof is not yet established for mechanically ideal plan recognition in visual perception.

The NEXT component was constructed in the expectation that the plan hypothesis would succeed and the belief hypothesis would fail. This predicts a pattern of low probability actions followed by a higher probability action through the joint probabilities of two action patterns which reflect different perceptual organizations: (a) schemata for all the actions expressed individually, and (b) schemata which link the actions by the NEXT component but thereby register only the point of action definition at the last action in the sequence. The conceptualizations expressed by the two patterns would be different but related by the **IN-ORDER-TO** relationship. For example, the set of actions which can be phrased "the small triangle moved to the door, he opened the door, and then the small circle escaped" would have the alternative plan organization roughly phrased "the small triangle moved to the door in order to open it and release the small circle." However, with the relative failure of the plan recognition hypothesis, we cannot yet rely on the NEXT component as a mechanism for structuring the perception process. In the computer analysis of the film this component was relegated to the less compelling role of handling curving trajectories for

[10]Not enough instances to warrant squaring for quadratic.

actors, as in the action "the small triangle and circle went around the box". The use here was dictated more by the choice of first-order change descriptors than theoretical motivations.

Future Research

Cartoon animation systems.

Cartoon animation systems permit an animator to create simple animations in two or three dimensional space for computer analysis and for real-time play-back. Such systems have been developed for AI studies of cartoon generation (Kahn , 1977, 1979). Of more importance than graphic realism (which seems to be the current trend) is that the system create long films which tell visually compelling stories.

The major advantage of a computer system for playing back the film is that the physical events can be investigated in depth using action schemas or, perhaps, more complete versions of the system for artificial perception outlined in this paper. Another major advantage is that cartoon animations are already digitilized, whereas other schemes require an extensive, and expensive, commitment of human labor.

Natural environment systems

There can be no question that cartoon animations impose a severe handicap on the quality of the study of action perception. Similar difficulties are apparent in other domains of computer vision and understanding. Animations provide action illusions, just like drawings provide object illusions. Newtson's work, in contrast to the present work, deals with records of natural events (except for the one study with the Heider and Simmel film (Massad, Hubbard and Newtson, 1979). State descriptions and first-order change descriptions for human and animal motions can borrow from various movement notation schemes (Newtson, 1976; Newtson and Engquist, 1976). These characterize continuous movements (first-order changes) as rotations about joints and the like. However, scoring films for such movements is extremely tedious and time consuming. When matters turn to computing first-order changes in natural environments automatically, the definitions of these changes become much more complex than the definitions permitted with cartoon animations.

Ultimately, an action recognition machine should be used on natural imagery, and not be dependent on cartoons. The problem is that this approach appears to demand that we wait until work in object and motion perception is completed to provide us with the machines which will interpret the objects and motions. In the animation system this is not a problem because the films can be completely encoded, objects pre-labeled, and motions defined using simple schemes. The closest we are likely to come to mundane situations where experience in action perception for natural human actions can be gained appears to be surveillance situations. The required characteristics are (a) fixed (unobtrusive, fixed perspective) camera position and (b) well-defined and limited domains of application.

The "image processor" for computing state and first-order change information for surveillance systems also suggests how useful but incomplete object and motion information can be obtained in natural environments. This processor capitalizes on fixed camera position by permitting hand-segmentation and hand conceptual-labeling of the background scene along with manual input of perspective criteria. The state description processor computes segmented occlusions of the background as possible un-labeled objects. The positions of the un-labeled objects can be described relative to the positions of known, conceptually labeled, objects (as in, "an object is near the door"). Movement and direction of occlusion boundaries could provide much of the first-order change information. Movement detected within the occluding object by time-down gradient operators may provide for further segmentation. I envisage that the output of this image processor is a partial description similar to the one generated for our experimental study.

It is easy see how to fool such a processor, so it is important to select surveillance situations carefully. Where there is informed deceptive action the surveillance devices would need to be "less smart", at least originally (see the discussion on "primacy effects"), but some forms of safety-related, unobtrusive, or deception-hardened surveillance remain interesting. Such contexts motivate further exploration of the expectation mechanism in its role in shifting and focusing attention for verbal description. Of course, for better or worse, such talk of exploiting real world circumstances leads to talk of "HAL 9000" computers (from the movie "2001: A Space Odyssey" c. 1969) that keep watch on peoples' activities.

Language and Perception Mappings

The relationships between language and perception have been under discussion for many years. A most famous hypothesis about this relationship is the Whorfian Hypothesis (Whorf, 1956);[11] that *language provides the structure to perception*. In present terms, the CON's available for action schemas are constrained to be CON's derived for simple English (not necessarily Russian or Hopi), and their presuppositions and entailments constrain the COLLECT and ACHIEVE propositions. We have already seen how logico-linguistic analysis of English expressions relates to action schema derivations. More experience is required with the present paradigm to properly evaluate the status of the Whorfian Hypothesis within it.

One idea for developing a system for artificial perception of actions would be to develop a brute force Whorfian perceptual machine. This has an extreme advantage over some alternatives: The mapping to English is planned into the design of action recognition schema. It is wise to anticipate the problem of English description because the mapping is so complex (see Waltz, 1980) , for additional insights). An added advantage is, perhaps, that the burden on what the system should see can be placed on the users of the system. The users describe, in English, what the system should see, and the system is then engineered to see it. This tactic has worked well in the implementation of practical systems for natural language understanding.

The serious fallacy in such a Whorfian perceptual machine is that presuppositions and entailments of English expressions may over-specify action schemas. This was the case in the protocol study. In a two-dimensional world where surfaces are not hidden from view, this is surprising. In a three-dimensional world view, inferences would have to carry part of the load in the computation. The expectation mechanism is one reasonable place to adjust over-specified action schemas against the realities of perspective viewing.

Object and Motion Perception

In this study I have de-emphasized work in object and motion perception (and their kin) in order to emphasize action perception. I noted that *interpreting* objects and motions may not be required in interpreting actions. The general idea is that environmental cues may be used in perceiving actions that permit hypotheses about the appropriate conceptualizations for objects and motions.

---

[11]Also called the Sapir-Whorf Hypothesis (Leech, 1974), linguistic determinism, and other things.

Object and motion conceptualizations could be elicited not only by bottom up cues but by recognized actions as well. This idea is very much in the spirit of recent work on motion perception which seeks to show how object recognition can be facilitated by motion computations (Ullman, 1979). The action schema mechanism can, in principle, resolve ambiguities about objects and motions (after actions are recognized), but it remains to be seen whether this avenue is worth exploring.

A more serious concern arises in the assumptions made about the computations of state and first-order change descriptions. Basically, I have assumed that these computations are carried out independently of the computations for action schemata. The only defense for this assumption is that it is often reasonable, and (b) it certainly makes it easier to gain experience with the physical antecedents of action perception. But then, the present approach is clearly deficient in dealing with the perception of actions in static forms (Birdwhistell, 1970; Herman, 1980) .

This paper began with the thesis that actions are perceived "directly", yet the *action postulate* asserts that actions are defined in second-order changes. The issue is whether the account of how actions are perceived need reference to how motions and objects are perceived. In the present formulation, motions are derivative in first-order change descriptions but they are not the same as first-order change descriptions. For an action to be derivative in second-order change, which is then derivative in first-order change does not imply that we have studied motions. We have not developed a theory of selection in first-order change, only one for second order change. If this strategy is fruitful, then the meaning of "direct perception of actions" is preserved. Action, motion, and object perception are different phenomena in terms of their computation.

A Derivative Scheme for Robotics

An interesting research domain is the artificial perception of actions by mobile robots. This domain would permit one to exploit the derivative or indirect aspect of action perception. The necessity of a fixed camera position is relaxed because the robot navigation system can provide object knowledge and perspective knowledge to the image processor. Unfortunately, it is much less practical to hand segment the scene for its stable (unchanging) parts as proposed for surveillance systems, since this would have to be done from many perspectives.

However, robot navigation schemes are now being designed that navigate the robot through the use of internal world models. In effect, the robot vision (and sensing) system maintains an internal model of the external world. Having designed the artificial perception of actions to work with derivative, state and first-order change information allows us to consider action perception from the internal world models. This strategy should facilitate the implementation of action perception in robots.

1. Acknowledgement

I had direct knowledge of Newtson's work over several years, and this paper represents a brief, nine-month, opportunity I had to work directly with him and some of the data he had collected over the years. I would like to thank Darren Newtson and John Rotondo for an enjoyable few months collaborating on a topic we all find intriguing. The work has been done, on the sly, with internal funding from The University of Virginia and the Carnegie-Mellon University Robotics Institute.

References

I. Appendix: State and First-Order Change Language

This describe the state and first-order change finite language composed for the Heider and Simmel (6) film. All numeric information is implicit in the symbolic descriptors and all assignments are deterministic. The language is not arbitrary and reflects heavily on the protocol analysis and the frame-by-frame analysis of the geometric and perceptual idiosyncrasies of the film. It is therefore particular to the film. It is included here for completeness.

The objects, O, in the film are a small circle (SC), small triangle (ST), large triangle (LT), a box (BX) and a door into it (DR). The variable P stands for any of the persons, SC, ST, or LT. The variable I stands for any of the inanimate objects, BX or DR. The variable R stands for any region as seen in Figure I-1. Figure I-1 also shows labeled containment boundaries (defined below), the unit of measurement, and the actors to scale. The following propositions are defined:

(Touch P O) An irreflexive proposition recorded when any line segment of the actor, P, physically touches any line segment of the object, O.

(Break I) An object, I, is broken when the line segments of I are broken.

(Rotate P) ST or LT rotated in either direction to any degree.

(Line-of-sight P P) An irreflexive proposition recorded when a line drawn between the centroids of the two P's does not intersect any line segment of any I. (This seems perceptually correct for the film.)

(Door Open/Closed/Opening/Closing)
Closed requires that the outer end of the door be touching the box; the definitions of Open, Opening, and Closing are transparent.

(At P P) Symmetric and Irreflexive: A actor, P, is at the location of another P if they are within 7 units of each other's centroids and there is a line-of-sight between them.

(At P R) A actor, P, is at the location, R, if his centroid lies in an area identified in Figure I-1, where R is offscreen (OF), outside the box (OT), inside the box (IN), the wall (WL), the doorway (DW), the door (DR), a lower-left region (LL), a bottom region (BT), a lower-right region (LR), a back region (BK), an upper-right region (UR), and a top region (TP). Note, by implication, if an actor is both IN and LR this places him in the lower-right corner inside the box.

(Near P P) A symmetric and irreflexive relation between two actors if their centroids are within 14 units and there is a line-of-sight between them and they are not at each others location.

(Move P Slow/Normal/Fast)
Translation of centroid by thresholding at 1.5 and 4.0 units.

(Moveto,MovePast,MoveFrom P P)
For two actors, P, there is an asymmetric, irreflexive relation: if there is a line of sight one actor will move toward, past, or from the other. The method is to determine the vector angle between the direction of movement and the direction of the other actor with the reference in the moment before the present moment (law of cosines). The angles are in 120^{o} arcs.

(Moveto,MoveFrom P R)
For an actor and a location, R, there is movement toward and from but not past. Movement possibilities are provided as labeled containment boundaries (double arrows) in Figure 1: offscreen-to-left (OFLF), offscreen-to-top (OFTP), offscreen-to-bottom (OFBT), offscreen-to-back (OFBK), outside (OT), inside (IN), wall (WL), door (DR), lower-left (LL), lower-right (LR), upper-right (UR), top (TP), back (BK), and bottom (BT). The method determines the nearest labeled containment boundary intersected by the directed movement vector of the actor. As in directional movement

between actors, where an actor was in the last frame determines where he is seen to be moving to or from in the present frame. The moveto vector is computed as if he was at the last frame location and the movefrom vector is computed as if he is at the present frame location. To illustrate, if an actor is above the box and moves rightward his movement is interpreted as going to the BK, not OFBK or UR. If he moves left, his movement is interpreted as going OFLF.

Figure I-1: Film Frame Worksheet for Appendix I.

Artificial Perception of Actions

Robert Thibadeau
Robotics Institute
Carnegie-Mellon University
Pittsburgh, PA 15213

Revision for *Cognitive Science*
17 February 1994

Address and phone: Robert Thibadeau, Robotics Institute, Carnegie-Mellon University, Pittsburgh, PA 15213, (412) 578-3824

Abstract

This paper has to do with the visual perception of actions that are discretely conceptualized. The intent is to develop a vision system that produces causal or intentional descriptions of actions, thus providing the conceptual underpinnings of natural langu age descriptions. The computational theory is developed in linking a "point of action definition" analysis to an analysis of how the physical events will elicit appropriate verbal descriptions. Out of this theory of direct computational linkages between physical events, points of action definition, and verbal descriptions, comes a theory of perception that provides some insight into how to go about constructing systems that can watch the world and report on what they are watching.

The aim of this research is to develop a machine that interprets visually observed events as discrete actions. Every discrete action is a momentary causal or intentional description of movement. To insure the system's veracity and usefulness, the description and thereby the "perception of action" must provide a faithful mapping to natural language. For example, the machine might verbally report that some person *is now walking toward the robot*, that a person *has left his station hurriedly*, that *a machine operator is pushing the wrong switch*, or that a person *has been walking down the corridor trying to open locked doors*, or that someone *is running* down the street. In effect, the machine must see everyday actions in a way that permits their ready communication to us.

There is hardly any active research on the problem of computing such perceptions. On the contrary, it is particularly easy to find papers that assert or strongly imply that the study of action perception must wait until the results are in on object perception and motion perception. This view is even expressed by researchers who will admit that motion perception can be studied without solving the problems of object perception.

It is my thesis that actions are perceived directly without necessarily interpreting objects or motions. I would like to understand how to account computationally for action perception without also accounting for all its ostensible precursors, like object and motion perception. Just as, in a computational sense, a motion can be perceived without first recognizing the moving object, it is possible an action can be perceived without first recognizing the underlying motions. Such a relationship should not be taken to deny indirect computation from object and motion *information*, only from the conceptualization of objects and motions. The present strategy approaches action perception intensely "bottom up" and "top down", seeking always to discover the most direct relationships that hold between the physical events and the verbal conceptions. To begin the analysis, we need to distinguish what is conventionally meant by the study of "action perception" by comparison to the studies of related perceptual phenomena such as "motion perception", "the perception of causality," and others.

Action Perception as a Study

Michotte's studies of the perception of causality (Michotte, 1963) represent some of the best known and earliest work of relevance to action perception. Michotte devoted his life to classifying physical conditions that elicit perceived causation. He discovered that perceived causality is based upon a few distinct sensations. There are only a few ways to perceive causation. By contrast, action perception is, by convention, not about distinct sensations of action, but about perceiving individual actions. The perception of causality is important in the perception of many actions, but it may not be involved in the perception of all actions.

Motion perception is often believed to encompass action perception, but the literature on motion perception deals with the perception of physical motions which are not inherently discrete. The physical motions are algebraically characterized by a cyclic or, at best, continuous, function with no naturalf beginning or ending (see Johannson, 1975; Restle, 1979; Ullman, 1979 and below). A classic problem, described in Ullman (1979) , is to understand the computation involved in merging spatially separated motions together into the perception of a single coordinated motion (e.g., seeing an arm movement as a coordinated movement of one object). Motion perception and object perception then become closely related topics. By contrast, action perception is further removed from object perception, *per se*, and is involved with discrete motions of objects. In action perception, the emphasis is as much in the conditions for the beginning and ending of motions as it is with the motions themselves.

One area of research associated with action perception is natural language understanding. A typical problem in language understanding is the understanding of natural language descriptions of actions at a "conceptual" level. There is relevant work in basic action categories (*primitives*, see Schank, 1975) and (Miller and Johnson-Laird, 1976) in which the aim is to compose complex actions out of simpler, more primitive, ones. There is also relevant natural language work on recognizing the significance of action sequences (Schank and Abelson, 1977; Schmidt, Sridharan and Goodson, 1978). The work in recognizing the significance of action sequences is of importance owing to the "hierarchical" nature of action. An action, although perceived as a single thing, is typically made up of finer actions concatenated together (e.g., body movements are a component of walking). The finer-actions can be individually perceived if the perceiver has the opportunity, and wishes, to see them.

The work from natural language understanding claims to describe a "language-free" conceptual level of representation which, by the claim, should bear on the understanding of action perception. Unfortunately, it is not that clear that "language-free" levels of representation have yet been described. Present knowledge about conceptual representation *de facto* derives from research grounded in the uses of natural language. It is true that the peripheral taxonomic structures of language, such as syntax and morphology, are distinguished from conceptual, or ideational, structures, but, from the point of view of the present research, what a linguist, or even a logician, may call ideational structures look, because of the derivation, like language structures as well. The same may be said for "conceptual representations" or "propositional representations" in AI. A major goal of the present research is to provide some independently motivated mapping or derivation between visual perception and our understanding of conceptual representation.

The study of "action perception" has been most developed by social psychologists, most prominently Heider (Heider,1958; Heider and Simmel, 1944) although I focus on Newtson (Massad, Hubbard and Newtson, 1979; Newtson, 1973; Newtson, 1976; Newtson and Engquist, 1976; Newtson, Engquist and Bois, 1977; Newtson and Rindner, 1979; Newtson, Rindner, Miller and LaCross, 1978; Wilder, 1978, 1978) . The interest in action perception stems from a struggle to understand how people attribute intention and causation to movements. Newtson's work is of particular interest to me because it provides an empirical, objective, and observational means of coming to terms with the process of perceiving actions. In his basic paradigm, informants watch an action sequence and indicate where they perceive actions as having occurred. Specifically, they are asked to push a button every time that, in their opinion, an action occurs in the film or video-tape they watch. The researcher records the times of the button presses for later examination with regard to the activities in the film. Newtson's "point of action definition" provides the impetus for my study of the artificial perception of actions. The premise in using a psychological basis is that people perceive actions in a fashion desired of the machine.

Points of action definition provide more than simply an idea about how others perceive actions. Most important, they provide timely clues about an ongoing process of perceiving. The alternatives, verbal reports and subjective analysis, are either not timely or just immensely difficult. A point of action definition is a simple and direct report from a person about *when* the action is seen. It turns out that the explicit point of action definition associated with verbal report and the observed physical event reveals critical relationships between the percept and the physical event. An understanding of these temporal and substantive relations is the basis for a computational understanding.

Newtson's conclusions about action perception were acquired in over ten years of research. I cannot summarize all of that research here. But one particularly relevant result is that people normally segment events into actions, even when they are not required to do so by instruction. However, people are remarkably unaware of their segmentations: A commonplace example is in film montage effects: cutting from shot to shot without confusing or making the viewer aware of the cuts. Film editors know that cuts must be made at points of action definition.

A final relevant conclusion from the research findings, is that action perception is susceptible to strong forces of expectation: viewers must be prepared to see an action in order for them to see it. The nature of this preparation is not well understood, but it is clear that a preparation or expectation system of significant conceptual impact operates in the normal perception of actions.

"Actions" in Visual Perception

This section focuses on the general problem of deriving knowledge representations of perceived actions. It begins with a conventional logical and linguistic (logico-linguistic) derivation and then develops a perceptual derivation of knowledge representations to take account of bottom up cues from the visual stream.

Top Down

There are many schemes for taxonomies of conceptual actions which attempt to simplify the representation of actions for conceptual understanding and inferencing purposes. While most all existing schemes pay little attention to how actions might be recognized perceptually, many carefully attend to the conceptual representation of actions. Schank's (1975) work with primitives in narrative prose is a familiar example and Miller and Johnson-Laird (1976) provides another. The work by Miller and Johnson-Laird is compelling because, although their technique was based on logico-linguistic analysis, their aim was to be accountable to visual perception as well as language perception.[12] To see how their logico-linguistic analysis works, and how it might work for visual perception, we can examine their analysis of some primitive actions.

The most primitive action in my system is loosely verbalized as "travel." This action was tentatively assigned the definition provided formally in (1). The notation has three parts: the proposition, **TRAVEL(x**), its presuppositions[13] (such as the presupposition that there is a place *y*), and its entailments

---

[12]Miller and Johnson-Laird (1976) provide an informal computational framework, but that computational framework will not be discussed in this paper. Their formal analysis of natural language and conceptual structure is, nevertheless, a classic contribution to our knowledge for its methodology, its depth, and its scope.

[13]"According to Strawson (1952) , statement *S* semantically presupposes statement *S'* if *S'* is a necessary condition for the truth or falsity of *S*." Page 172 in Miller and Johnson-Laird (1976), suggests a willingness to accept a pragmatic or interpersonal version of this definition which depends on belief states.

[14](as in (i)). The time indicator t is conventional. Miller and Johnson-Laird's use of the *statement forming operator R* (Rescher and Urquhart, 1971) is possibly novel: R_t corresponds to "this statement is recognized as true at time t". Another statement forming operator, Q_t, is used later and it means that "this statement is recognized as true at all moments prior to t".

(1) **TRAVEL**(x): Something x "travels" if there is a time t and a place y such that:

(i) Either: R_{t-i}(not**AT**(x,y)) & R_t(**AT**(x,y))

Or: R_{t-i}(**AT**(x,y)) & R_t(not**AT**(x,y))

The formulation in (1) pertains to conceptual representation, but I believe it superior to other conceptual formulations, such as Schank's PTRANS, in that it stipulates logical presuppositions and entailments for primitive actions. I will later evaluate a method of action recognition that seeks to match such semantic "necessities" against visual information. The proposition, **TRAVEL**(x), nevertheless, lacks the place descriptor y, suggesting its natural language or conceptual function. In reasoning and understanding, the particular place from which or to which the object moves is often not relevant. Miller and Johnson-Laird (1976) state through presupposition and entailment how one may elicit the perception of an action.

The statement forming operator may tell us when the conceptualization is recognized. If the conceptualization is an action conceptualization recognized at a single point in time, then we have a natural way of predicting a point of action definition. But, in fact, statement forming operators account for verb tense and aspect in Miller and Johnson-Laird (1976) and may only accidently predict points of action definition. This first version of **TRAVEL** has problems with the "point of action definition" property. The statement forming constraints specify two points of recognition, R_t and R_{t-i}. Furthermore, the entailment is *ambiguous* in point of definition: the point of definition is either the moment when the object moves from a place or the moment when the object arrives at a place.

For logico-linguistic reasons, Miller and Johnson-Laird were not satisfied with this first definition either and went on to develop their final formulation in (1'). They wanted to derive that "if x does not travel, it stays where it was." They also felt **TRAVEL** should have a durative entailment. They did not want to allow, as (1) seems to imply, that **TRAVEL** is defined when objects just disappear or appear at locations.

(1') **TRAVEL**(x): Something x "travels" from time t_0

to time t_m if, for each t_i such that

$t_0 <= t_i <= t_m$, there is a place y_i such that

R_{t_i}(**AT**(x,y_i)) and:

(i) $R_{t_{i+1}}$(not**AT**(x,y_i))

[14]At first sight, entailments are simple conditionals: a statement S entails statement S' when S is a condition for S'. When S is true, S' cannot be false, but when S is false, S' can be true or false. However, in Miller and Johnson-Laird's usage, the negation of S necessitates the negation of S' (see page 532). By example, if it is true that "x travels" then *(1)(i)* holds, if it is true that "x does not travel" then the negation of (i) must hold as well. If either antecedent, "it is true that ..", had been false, then the consequent could have been true or false. This is a proper usage of entailment, and I will adopt it in this paper.

TRAVEL$_{(1')}$ represents a simple motion which is, conceptually, the simplest form of action. But this is still ambiguous in point of action definition because, by continuously changing the focal place, y with y_i, it now describes a primitive motion (with no fixed beginning or ending) and thereby no point of definition or, perhaps, an entirely filled sequence of them. Later I will show a way to resolve this problem and derive the primitive action from the primitive motion described in **TRAVEL**$_{(1')}$.

The sense of *primitive action* is that actions of greater complexity are built of Boolean predications which directly or derivatively employ the primitive predicate. The predicate **MOVE** from the English transitive verb "move" is the primitive **TRAVEL** made complex by a perception of agentive **CAUSE**. The verb "move" also has an intransitive usage, as in "he moves", which is not addressed here. Rather, we have:

(2) **MOVE**((x),y): Something x "moves" something y if:

(i) **TRAVEL**(y)
(ii) **DO**(x,**S**)
(iii) **CAUSE**(**S**,(i))

The **DO** is, in Miller and Johnson-Laird (1976) and the views of many others, a placeholder for other movements, This placeholding is not without presumed consequence. It is a further manifestation of action "hierarchies," in this case a form of one action (the **TRAVEL**) subsuming whatever action(s) composed the **DO**.

The relationship between the **DO** and the **TRAVEL** is potentially more complex than a **CAUSE** (Michotte, 1963 can be consulted for an appropriate perceptual analysis of causation). If the action is perceived as intentional, as this sort usually is, then the subsumptive relation is an **IN-ORDER-TO** relation: the actor did something in order to move the object. The **IN-ORDER-TO** relation has been worked out by analysis similar to Miller and Johnson-Laird (1976) in Schmidt and Sridharan and Goodson (1978).

The added entailments in (2) also add complexity to the question of resolving points of action definition. Causation, for example, is perceived at a moment in time and could predict a point of action definition. However, other complex actions establish points of action definition in explicit ways and without causation. Consider the presuppositions and entailments for "reach" and "depart" in (3) and (4). The actions are momentary, at t, and each action presupposes a past, durative, activity as indicated by the statement forming operator Q_t.

(3) **REACH**(x,w): Something x "reaches" some place
w if there is a moment t such that
Q_t((**TOWARD**(**TRAVEL**))(x,w)) and:
(i) R_t(**AT**(x,w))

(4) **DEPART**(x,w): Something x "departs" some place
w if there is a moment t such that
Q_t(not**TRAVEL**(x)) and:
(i) R_t(**TRAVEL**(x))

With these formulations, the referential ambiguity in **TRAVEL** is reduced by additional statement forming and presuppositional constraint. The point of action definition is suggested as the moment, t, at which the actor reaches a location or begins to travel.

The illustrations with the "move," "depart," and "arrive" demonstrate a taxonomic system for action we would like to have available to processes of action perception. However, there is a failure in the logico-linguistic analysis to support such a scheme for perception. We must look elsewhere for that analysis.

Bottom Up

Another method for approaching action definition carries us further from natural language considerations and closer to deriving conceptualizations from physical events. This previously unpublished definition was formulated by a colleague, John Rotondo, working with Newtson to develop mathematical (not computational) models. To set the stage for this method we make a simplifying assumption that it is possible to describe the effective input into the perceptual system as a sequence of states, **S**; where a state, s_t, is an interpreted image[15] at a moment, t, in time.

The generative method recognizes two classes of change in a succession of states, **S** at s_t: a simple state difference or first-order change, c_t^1, and a change in a change or second-order (derivative) change, c_t^2. The number of moments contributing to these changes is not specified: successive state differences can compute a constant curving movement, a uniform acceleration, or constant cycle, as well as straight movements and constant velocities. A change in a change can just as well be a change in direction as a change from movement to no movement.

A generative description of the domain of possible descriptions of an event discretely continuous over time is obtained by evaluating all change descriptors for all moments in time. This generative description, it is claimed, *generates the superset of all actions in an activity*, as well as all motions. It takes little imagination to recognize that the number of generated actions and motions rapidly grows very large. Even with dramatic simplification, such as differencing between just two successive states, real world imagery produces enormous numbers of change descriptions over short periods of time. We can, however, now view action perception as problem solving and formulate action schema in terms of a search method through a space of descriptions.

Three reductions are used to limit the search:

- Points of action definition are drawn from the collection of second-order changes, $\mathbf{C}^2$. This is called the *action postulate*.
- A *systemization of schemas* which recognize actions is provided. This is similar in spirit and form to Miller and Johnson-Laird (1976) .
- A system for dynamically elaborating schema and making selected schema available to perceptual processes is provided called the *expectation mechanism*.

The action postulate. The action postulate is advantageous when it is important to monitor changes, such as a continuous movement toward a place, over some period of time, just as one monitors states, such as an object at rest. In describing the perceptual activity, as opposed to the conceptual activity, monitoring environmental invariances becomes important (Gibson, 1966).

[15]Two, two and a half, or three dimensional segmented, and perhaps conceptually labeled, representation of the visually encompassed world for a moment in time.

The *action postulate* was derived empirically.[16] used the term "feature of change" rather than "second-order change" in his description of the postulate.). Newtson's studies (some of which are not published) have suggested that the action postulate, while correct in one sense, is only approximately precise. By example: In a film of a man hammering a nail into a block of wood, people routinely report the action slightly after the second-order change has occurred. People differ in selectivity: some will report actions only when the man hits the nail while others will see the man raising his hand as an action as well. Nevertheless, they are consistent in reporting the action several instants *after* the hammer stops on the nail or *after* the hand is at its full altitude and just begins to fall. My own observations have confirmed such a delay. The only work with the delay has used film speed variations to confirm that the delay is not simply reaction time delay: As film speed is increased, the delay is shortened, and, as it slows, the delay lengthens. Unfortunately no systematic investigation of this delay for different classes of action has been undertaken to date.

In the next section, we develop a system to represent the action postulate explicitly and account in general terms, for the observed imprecision in the postulate. The system derives particular action schema from a parent action schema in a fashion analogous to the systemization in Miller and Johnson-Laird.

Action schemata for perception. The point in time where an action occurs provides concrete references for the undertaking in artificial perception, but as a practical computational matter, it is good to know how the machine is enabled to see actions defined at particular moments in time, from bottom-up cues. The cues I currently allow are the state descriptions, **S**, and first-order change descriptions, $\mathbf{C}^1$, available to perception.[17] The major part of the computational study was in how to define action schemata which would map these bottom up cues to action conceptionalizations.

I initiated the computational study without any presumption of a single primitive schema for actions. In fact, I set out to develop any number of elaborate finite state machines consistent with Newtson's observations that could be employed to recognize different actions. These state machines were found to be largely compositions of a simpler machine which I now think of as a primitive or parent action schema. This parent action schema and its systemization is the topic of this section.

The parent schema is defined as an automaton, thereby defining the schema in procedural terms, in the same spirit as Piaget's, (1963) formulations . The systemization asserts that *every action perceived is a manifestation of a modified instance of the parent action schema. In that manifestation, this instance has been evaluated against the sensory stream and has thereby successfully completed the instantiation of its associated action conceptualization*.

The parent action schema has five parts which includes the declarative conceptualization which the automaton functions to instantiate and assert, two special conditions it monitors in the sensory stream, a procedural method for robustness against noise, and a procedural subsumption method for forming compositions of schemas which would signify a single conceptualization. These components, their modification and evaluation, are described in detail below:

[16] Newtson (1976

[17] For this paper, I set aside discussion of recursively allowing second-order change descriptions to influence the computation of new second-order change descriptions, as in "The man repeatedly went to get coffee".

- *A Conceptualization*,CON, to stand for the action at the moment it is conceptualized.

- *A Criterial Component*,COLLECT, which is a Boolean composition of simple state and first-order change descriptions to be detected in the sensory stream. This component is not the only referential component criterial to an action, but it is the only referential component criterial to the point of definition of the action. All actions will be defined at the point in time the criterial component goes from true to false by reference. For example, a person breaking a glass might include monitoring the fall of the glass toward the floor. When the glass ceases falling, or perhaps moving (within a limited pattern), the action is defined.

- *A Non-Criterial Component*,ACHIEVE, which is a Boolean composition of state and first-order change descriptions to be detected in the sensory stream. This non-criterial component is consulted before and at the moment of action definition to determine whether the action itself is true or false (whether the glass broke or bounced). While this component is referentially criterial (it determines the truth of a proposition on referential or deictic grounds), it is not criterial in identifying the point of definition of the action.

- *A Perceptual Error Correction Component* which determines the tolerance in evaluating the change in truth value of the Criterial Component. This component represents a system of processes which modulate perceptual decisions: For example, it would guarantee that a brief stop in movement (owing to any error or weakness of man or machine) would be routinely ignored and not signal a false point of action definition.[18]

- *A Motion Linking Component*, NEXT, which permits chaining action conceptualizations (viz., CON pointers) for "macro" definitions. This is an optional component. The presence of a NEXT predication can suspend a true-false decision through the ACHIEVE Component, and thereby it can suspend a point of action definition. An example of the use of the NEXT component can be found in one of two ways of viewing a man hammering a nail (see Newtson, Rindner, Miller and LaCross, 1978): (a) simple, one schema, actions such as the hammer hitting the nail or the hand rising, and (b) a sequence of the hand rising then the hammer coming down and hitting the nail. The sequenced perception may be of use in stages of becoming skilled

[18]The specific error correction system generally depends on the details of the implementation. In the implementation described later, the system associated two parameters with each action schema, (a) a minimum duration threshold for the COLLECT proposition to turn true (after verification begins) and (b) a maximum duration threshold over which the COLLECT proposition was not true before deciding the COLLECT proposition was in fact not true.

at perceiving. If the hand rising has no other effect than to allow it to fall, the action schema may be shortened to contain only the last part of the hammer hitting the nail. A further reason for the NEXT is that the processes which adjust what actions are "seen" will no longer posit "the hand rising" as a relevant thing to monitor once the NEXT relationship is recognized between the rising and the hitting.

Instances of action schema can be derived for Miller and Johnson-Laird's (1976) examples. I will use a one state proposition, **At**(x,y), which signals that an object, x, is at a place y, and one first-order change proposition, **Move**(x) which signals that an object has changed its position in space. To recognize a **TRAVEL**$_{(1)}$ requires at least two primitive instantiations of the parent schema to conform with the ambiguity in its entailment. The instantiations are given in (5) and (6). The new representation fills designated slots in the schema described above.

(5) CON: R_t(**TRAVEL**$_{\mathbf{(1)}}$(x))
COLLECT: **At**(x,y)
ACHIEVE: not**At**(x,y)

(6) CON: R_t(**TRAVEL**$_{\mathbf{(1)}}$(x))
COLLECT: not**At**(x,y)
ACHIEVE: **At**(x,y)

The action postulate says that the point in time, t, that the COLLECT for (5) or (6) goes from true to false, there is a determination of the truth or falsity of the conceptualization, CON. In these cases, the conceptualizations, if ever recognized, will be true, since the COLLECT is the inverse of the ACHIEVE. A more serious problem with this formulation is that it violates the action postulate since it only detects a first-order change (change between two states), not a second-order change (a change in a change). So, like Miller and Johnson-Laird (1967) we are led to reject this formulation as unsound. **TRAVEL**$_{(1)}$ conflicts with the action postulate.

Two different schemas are required to recognize the durative form of **TRAVEL**$_{(1')}$. It is not possible to report the ongoing change at an arbitrary moment: we must select the beginning of the travel as in (7) or the end of the travel, as in (8).

(7) CON: R_t(**TRAVEL**$_{\mathbf{(1')}}$(x))
COLLECT: not**Move**(x)
ACHIEVE: **Move**(x)

(8) CON: R_t(**TRAVEL**$_{\mathbf{(1')}}$(x))
COLLECT: **Move**(x)
ACHIEVE: **Move**(x)

These schemas collect the absence of a change of position or the change of position. When either of these conditions change (go from true to false), the system reports the object, x, has traveled. The apparently unambiguous **TRAVEL**$_{(1')}$ has two forms, like **TRAVEL**$_{(1)}$, but because of the imposition of a natural point of definition on the action of traveling. These, nevertheless, represent primitive conceptual actions by perceptual derivation. **TRAVEL**$_{(7)}$ is virtually the same in semantic form to **DEPART** in (4). However **TRAVEL**$_{(8)}$ corresponds to a past tense version of primitive movement **TRAVEL**$_{(1')}$.

The current system provides for the delay between the physical moment, t-i, when the COLLECT proposition goes from true to false, and the response moment *t*, when the point of definition occurs, in the simplest fashion possible. Because of the effect of noise, the perceptual error component allows some absorption in deciding whether the COLLECT proposition has gone from True to False. I assume that the same absorptive process that keeps the COLLECT proposition True in noise also accounts for the observed delay (perhaps as much as several seconds) in reporting that an action has occurred. Since the COLLECT monitoring rate is defined by reference to the environmental frame rate, not in absolute time, this explanation is also in agreement with the preliminary experimental findings when film projection rates are speeded or slowed.

Defining action schemas for recognition is clearly different from defining them for conceptual analysis, but the taxonomic principles of schema organization are potentially similar. Instantiations and compositions of the instantiations[19] of the parent schema provide the basis for taxonomic categorization of action schema. Since conceptualizations are a product of schema recognition, the taxonomies can naturally reflect the taxonomies of conceptualizations. But this sort of taxonomic constraint is definitional and not very interesting.

It would be interesting if we could also carry down the conceptual constraints (such as the "toward" modification on **TRAVEL** which contributes to **REACH** in (3)) to the COLLECT and ACHIEVE levels of appropriate action schemas. Formally, COLLECT carries presuppositions and ACHIEVE carries entailments: a True COLLECT proposition is required even to entertain the truth or falsity of the CON. But given that COLLECT is True, a True or False ACHIEVE proposition determines whether the CON is True or False. This is an important connection between the linguistic and perceptual analysis.

The match also depends upon the first-order change and state descriptors that result from bottom up processing. As a research tactic, these descriptors can be adjusted to guarantee compatibility. The most dominant factors in carrying out conceptual constraints are, then, the expectation mechanism and the presence of the required information in the physical events. These are the next topics.

Expectation Mechanism. The model of action perception requires the generation of pools of action schemas to act as independent finite state automata that monitor the state and first-order change information on a moment by moment basis. Action schemas monitor derivative information, not the physical events directly, in as much as their function is to respond to second-order changes.

The relationship between action schemas and state and first-order change information is fundamentally a matching relationship where current action schemas are repeatedly matched until the COLLECT proposition in any one goes from true to false and an action is defined. The *expectation mechanism* has the role of identifying the current subset of action schema which are, for one reason or another, worthy candidates for such matching. The expectation mechanism may both *generate* and *select* action schemas.

There are reasons to limit the set of current action schemas to a subset of action schemata.

- To do a physical implementation which operates in something close to real time, it is of course necessary to limit the set size.

---

[19]Using the NEXT mechanism.

- Action schemas are all instantiations and compositions of instantiations of a most general action schema. There is no requirement for complete instantiations. For example, there might be a general "open a door" schema that matches anyone opening a door. If two schemas are available which differ only in specificity, then only the least specific schema requires matching. I know of one empirical observation that supports the availability of incompletely instantiated action schemata: If something unexpected occurs in an activity, the temporal density of points of action definition increases abruptly (Newtson, 1973; Newtson and Rindner, 1979). This effect has been called a *primacy effect*; it is more like an reorientation effect. It suggests that people initially use more abstract action schema to pick up on the "micro-actions" in order to make initial sense of what is going on.

- There are logical constraints on schema relevance: an "open a door" schema need only be monitored when a door is in the picture. There needs to be a mechanism which obviates the need for a 'close the door' schema to be monitored when the door is already closed.

- In addition to the relevance of partial matching and logical constraints, perceptual confusions are a functional reason for selectivity. Preliminary empirical work suggests this functional basis may be a most significant one. Over the long term, a variety of context mechanisms are important to reduce the complexity of action schema. The fact that a COLLECT proposition represents any arbitrarily complex Boolean predication of state and first-order change descriptors does not assure that any arbitrarily complex Boolean predication of such descriptors is in fact sufficient to distinguish one action from another. Perception is selective in profound ways: the same activity and the same physical manifestation may have different significance on different occasions. It is well recognized in Epistemology that one can take virtually any action one can conceive of and make up an example: "closing the door" may alternatively be seen as "pushing the door as far as it will go" (with exactly the same schema except for the CON designation). The difference may be in some higher intention, in the example, containing something versus getting the door out of the way. The NEXT component was designed specifically for this reason. In the hammering example, raising the hammer then (NEXT) bringing it down seen as a single action by a perceptual sequence may, with further accommodation, be seen as hitting the table with the hammer. Also recall the **DO** in linguistic semantics: Raising the hammer and lowering it is what the actor **did in order to** hit the table. This problem has been extensively addressed from an AI perspective by Schmidt and Sridharan (Schmidt, Sridharan and Goodson, 1978; Sridharan and Schmidt, 1978). I have, in effect, adopted their resolution which posits the use of an expectation mechanism.

A reasonable framework for talking about an expectation mechanism is in terms of a generalized rule system, or production system (Thibadeau and Just, 1982). The function of the rules in the production system is to adjust the expectation set. The only adjustment which can be made is the addition or subtraction of action schemas to and from the expectation set.

Within this framework, I propose a new postulate, the *expectation postulate*: *the expectation set is adjusted only at points of action definition*. The rules in the production system add or delete action schemas from the expectation set in response to actions. One of the reasons for identifying a production system as a useful preliminary model for an expectation generator is that it is structured to respond quickly to new demands (Erman and Lesser, 1978; Tanimoto, 1982). The information available to the rules in the production system can be considerably richer than simply the most current action recognized and the set of all action schemas. Furthermore, the rules can posit intermediate results and can apply with intermediate effects other than simply the addition and subtraction of action schemas for the action recognition matcher.

A potential practical problem with the expectation postulate is that artificial perception could be effectively blind to unexpected actions at least for a significant period of time. To reduce blindness, a proper role of the action recognition matcher would be to carry an implicit action schema that can signal significant activity not being accounted for by any action schema. This could be given by an action such as $\mathbf{TRAVEL}_{(7)}$(x) signaled at the beginning of a movement with a free variable for x.

From one perspective, the action $\mathbf{TRAVEL}_{(7)}$(x), once recognized, is treated like any other recognized action: the rule system provides additions and subtractions of action schema in response to the action. It is likely that having the "wired in" accounting mechanism alone is more desirable than a "wired in" $\mathbf{TRAVEL}_{(7)}$(x) schema. An explicit loading of situationally sensitive $\mathbf{TRAVEL}_{(7)}$-type schema is preferable. Such schemas do not respond unless the second-order change they detect is not accounted for in any other action schema (they always compute off residual state and first-order change descriptions).

A Complete Model

Having a parent action schema and an expectation mechanism does not tell us the content of schema instantiations and how to compose them. Logico-linguistic methods seem appropriate for filling out conceptual representations. But a different method is called for in action perception since the problem is one of defining the perceptual engine, not manipulating the content it generates. I believe that empirically oriented "protocol studies" are the most useful: informants provide not only their English descriptions of the actions, but also an indication of points of action definition. Action schemas can then be tailored according to human demands in a complete computational model which goes from raw visual imagery (e.g., Cartesian data or pixel data) to conceptual description.

I had an opportunity to do one such analysis which included a protocol analysis and experiment, and the results of that study are promising. An old animated film was selected for the study. The film was constructed on a "flatland" theme by Heider and Simmel (1944) . The film has been extensively discussed, and peoples' perceptions of it extensively researched, at least since 1944 (Greenburg and Strickland, 1973; Heider and Simmel, 1944; Massad, Hubbard, and Newtson, 1979; Shor, 1957.) The story-line of the film is not trivial, although the characters are two triangles, a circle, and a box with a hinged door on it. Appendix I has a picture). Let me reconstruct the story in English terms (since it is not possible to faithfully render a movie in a page of print). In this reconstruction, I have tried to be faithful, in flavor, to how one might first view the film (for example, there is a vivid impression of the gender of the actors, which is retained in the text):

> A large triangle is just inside the box at the open door. He closes the door and goes to the bottom of the box.
>
> A small triangle and a circle enter the screen together. The small circle stops out front of the box while the small triangle heads around until he is at the door.
>
> The large triangle then goes to the door, opens it, and, almost immediately, he viciously attacks the small triangle. The small triangle is repeatedly hit against the wall of the box and finally goes under the box in apparent submission.
>
> The large triangle now goes after the small circle who has hidden in the box and closed the door. The large triangle gets into the box and closes the door behind him. He now tries to corner the small circle but she never gets hit or pinned down. Rather, the small triangle opens the door and the small circle escapes.
>
> The small triangle and small circle appear to kiss while the large triangle has difficultly trying to open the door which is apparently locked. The large triangle, on getting out of the box, chases the twosome around the box but they escape off screen.
>
> At this point, the large triangle moves over to the door, opens and closes it, and then hits it, breaking the door then the box into tiny pieces.

I asked three students to watch the film several times each. The result of the analysis of the protocols and the physical events in the film is a total of 64 action schemas with no variables for objects. The physical action schemas were applied with simple (logical and thematic) expectation adjustments to the film to yield a total of 149 instantiation points of action definition over the film.

Having the characters depicted as simple geometric forms in 2D space considerably simplifies many image analysis computations and allows us to study action perception rather directly. Note that although the movie shows only geometric shapes, the perception of sentient, thinking, characters is as vivid as in any cartoon animation. As a narrative, this cartoon is not at all simple. Actions by the actors are not perceived as simple movements on the screen, but as implication-rich intentional actions. There are examples of cooperation, vicious attack, submission, planning, surprise, and even "kissing". The story structure is a standard narrative form: the paragraph structure in the description correspond to the elements: (1) Setting , (2) Motivation, (3) Primary Engagement, (4) Primary Engagement, (5) Secondary Engagement, (6) Value Resolution.

In my computer analysis, I used the cartesian coordinates of all the line segments and arcs in every other frame of the 1690 frame film as the raw data about the physical events. A uniform computation of state descriptions and first-order change descriptions over the entire film was made from the Cartesian description (see Appendix I). The state and first-order change descriptions were generated by a FORTRAN program. A LISP program was used to analyse the theory of action perception. The protocol analysis involved comparing the human responses against the computed state and first-order change description, formulating action schemata to render those responses as faithfully as possible, and running the LISP program to confirm the results were reasonable. Of the nine months on this project, this work represented eight months.

The state and first-order change descriptions from the program for frames 538-542 are shown below. Frame 540 corresponds to the moment before the informants said that the large triangle intentionally hit the small triangle. Since we know a **HIT** was seen, the state and first-order change descriptors which first appear relevant have been printed in bold face. The rendition reflects the fact that the description at this level will omit a predication if it is referentially false: this was possible because we used the finite and deterministic referential language described in Appendix I.

(9)Computer generated State and First-order change descriptors for three successive frames.

Frame 538: At(<all actors>,outside-box)
At(small-triangle,wall)
At(small-circle,door)
Door(Open)
Near(small-triangle,large-triangle)
Line-of-sight(<mutual among all actors>)
Move(large-triangle,normal-speed)
Move(small-triangle,slow-speed)
MovePast(<large and small-triangle>,small-circle)
MoveToward(<large and small-triangle>,wall)
MoveFrom(<large and small-triangle>,left-border-of-screen)
MoveToward(large-triangle,small-triangle)
MoveFrom(small-triangle,large-triangle)

Frame 540: At(<all actors>,outside-box)
At(small-triangle,wall)
At(small-circle,door)
Door(Open)
At(small-triangle,large-triangle)
Line-of-sight(<large-triangle and small-circle>,small-triangle)
Move(large-triangle,normal-speed)
Move(small-triangle,normal-speed)
MovePast(<large and small-triangle>,small-circle)
MoveToward(<large and small-triangle>,wall)
MoveFrom(<large and small-triangle>,left-border-of-screen)
MoveToward(large-triangle,small-triangle)
MoveFrom(small-triangle,large-triangle)
Move(small-circle,slow-speed)
MoveToward(small-circle,<entrance-to-box and large-triangle>)
MoveFrom(small-circle,top-border-of-screen)

Frame 542: At(<all actors>,outside-box)
At(small-triangle,wall)
Touch(small-triangle,wall)
At(small-circle,door)
Door(open)
At(small-triangle,large-triangle)
Line-of-sight(large-triangle,<other two actors>)
<there is no movement>

It must be emphasized that the perception that the large triangle intentionally struck the small triangle is very strong despite the fact that the two triangles do not touch each other. Certainly *contact* is presupposed (and entailed) in hitting, yet the perceptual criteria appear less demanding. Note that actions are viewed from above, so physical interactions are not hidden from the viewer.

Other examples of hitting also suggested the need for weak, underspecified, definitions. The schema which correctly detects the 12 cases of hitting is shown in (10). Roughly speaking, the action is defined in a brief moment of apparent contact in which the aggressor is moving toward the aggressee.

(10) CON: **HIT**(aggressor,aggressee)
COLLECT: MoveToward(aggressor,aggressee) AND
At(aggressor,aggressee) AND
(MoveFaster(aggressor,aggressee) OR
notMoveToward(aggressee,aggressor))
ACHIEVE: At(aggressor,aggressee)

On another point, the Touch or contact predicate in Frame 542 was notated independently of the Cartesian description by the secretary who transcribed the film off the film editor. It was inconsistent of her to describe the small triangle touching the wall in frame 542 but not frame 540. This is simply a source of error which occurs since the various state and first-order change predicates can derive from different processes in the input stream. A recognition system that cannot routinely compensate against such minor errors is not interesting. Error in description certainly contributes to the lack of specificity in schema, but I confirmed that the film also lacked the necessary information.

As with the hitting schema, a number of other actions reported by the respondents were carried out more than once. A few of the other more commonly recognized action schema from tdhe computer analysis (discounting for actor instantiation as in (10)) are provided in (11-13). Again, the recognition criteria are weaker than might be supposed by an analysis of entailments and presuppositions:

(11) CON: **OPEN**(person,door)
COLLECT: Door(Opening) AND
At(person,door)
ACHIEVE: Door(Open)

(12) CON: **SHAKEorROLL**(person)[20] has reference:
COLLECT: Rotate(person) AND
notMove(person,normal) AND
notMove(person,fast)
ACHIEVE: Rotate(person)

(13) CON: **WENT**(person,door,(from inside box))
COLLECT: At(person,inside-box) AND
MoveToward(person,door) AND
notTouch(person,door)
ACHIEVE: At(person,door)

The actions in (11-13) are all distinct from the **HIT** in that they are durative. This provides for a small adjustment in the perceptual error component and that in turn agreed with the observed delays in the points of action definition. Furthermore, the point of definition for **HIT** is at the moment of causation, while the point of definition for the other actions is at a moment when a relevant movement ceases.

[20]The respondents used the language "shake or roll". So will I.

The translatory movement in (13) could have been generalized by eliminating the "from inside box" criterial feature, but there was a tradeoff between specificity and sheer numbers of points of action definition. A very general translatory schema provides many hundreds of points of action definition, even in this film. Nevertheless, what was compelling in the analyses was the fact that the actions predicted by the action schema matched reported points of action definition with precision good to approximately three Frames (approximately 1/8 second). The study leaves uncertain, however, whether it will be possible to describe action schemata in terms of presuppositions and entailments from corresponding English expressions. It does suggest that a unification of action schema taxonomies and action conceptualization taxonomies is hard to achieve.

Experiment: Plan Recognition or Local Intentionality?

Although the machine implementation included an explicit expectation mechanism, the protocols could not reveal much of its character. An experiment was undertaken to sort the 149 points of action definition in ways that might give us an indication of the nature of the expectation mechanism.

Action perception for intentional actions, the predominate class of action in the film, is often discussed as a function of plan recognition. Plan recognition has to do with specifying how to recognize the plans, beliefs, and goals of actors on the basis of their actions. The seminal AI work in this domain was done by Schmidt, Sridharan and Goodson (1972), Sridharan and Schmidt (1978), Schank and Abelson (1977), Bruce and Newman (1979), Solway and Riseman (1977), and Wilensky (1978, 1978). ,(Schmidt, Sridharan and Goodson, 1972; Sridharan and Schmidt, 1978, Schank and Abelson (1977) , Bruce and Newman (1979), Soloway and Riseman (1977) and Wilensky (1978,1978). The general supposition made by these researchers is that action perception is often explicable in terms of plan recognition. In other words, the actions attributed to an actor have to do with the beliefs one holds about the actors' intentions.

Schmidt and Sridharan noted the selective nature of action perception and developed elaborate schemes for adjusting expectations against knowledge developed in plan recognition. However, there has been no empirical assessment of such an expectation mechanism for action perception, *per se*. I assume that the case is made that plan recognition considerations are relevant to the expectation mechanism (further supposing an expectation mechanism is worth considering in the first place), and that there is interest in weighting the relative significance of different factors.

The 149 points of action definition generated by the protocol analysis represent a small percentage of all the actions that could have been specified, but it is fair to assume that the collection represents a sample biased toward the most likely actions consistent with the present theory. I sought to avoid further bias by testing two alternative hypotheses. These hypotheses adjust the probabilities of action perception differently for the qualitatively different actions. The points of action definition were those collected by Massad, Hubbard, and Newtson (1979) by 55 students watching the film for the first time. The reader should consult that article for the exact methods used in data collection. Newtson's studies have suggested that the only critical aspect of the method is that people are told to "push the button whenever, in your opinion, an action occurred." The experiment concerns the probability distribution of button presses over time by predicting probabilities of the actions reported by the machine in the protocol study. This was an experiment in the best sense: despite almost nine months of opportunity, during the machine implementation and protocol study, I did not look at the true probabilities of points of action definition until after the assignments of machine interpreted actions into hypotetical probability categories was made.

The alternative hypotheses both assume that actions are seen in terms of an actor's goals, but they weight probabilities of seeing an action by different kinds of goal significance. It is supposed that the probability of an action being reported is either

- due to the natural hierarchies of planning, viz., subgoal-goal structures for an actor's plans (*the plan hypothesis*), or
- the subjective certainty that the goal was the actor's goal, viz., the observer believes the actor has the goal (*the belief hypothesis*).

Neither of these hypotheses is highly quantifiable given present knowledge, but they do readily permit an assignment of the 149 actions into two or three probability classes.

The Plan Hypothesis. The higher-level plan structure of the film seems obvious and provides for competing goals of the actors. Shor(1957) confirmed this was true for 95% of the people who viewed this film.

1. The large triangle wants to beat up the small triangle and rape or beat up the small circle.
2. The small triangle and small circle want to prevent that.

Just to make sure, Newtson told the 55 people before viewing the film that it was about a bully and two innocent passerby. In any event, the major goals are evaluated through a *composition* of subgoal successes and failures in the film. The large triangle succeeded on occasion in beating up the small triangle and in at least successfully threatening the small circle. The small triangle and circle succeeded on occasion (and finally) in preventing this. A success for them was always a failure for the large triangle (except for a short episode of success in cooperative planning which dealt with the evident affection the small triangle and circle had between themselves). Specifically, the plan hypothesis generates three classes of action (in decreasing order of expected probability):

1. **LT-succeed**. The large triangle succeeded in achieving a subgoal which directly supports the supposition that his major goal is achieved. (For example, the large triangle's continued hitting of the small triangle conveys a sense of the large triangle succeeding with each hit.)
2. **ST&SC-succeed**. The small triangle and small circle achieve a subgoal directly supporting the supposition that their major is achieved. (For example, the release of the small circle from the box by the small triangle conveys a sense of the small guys winning.)
3. **Irrelevant**. The action does not directly support either supposition. Several of the door openings and closings are of this sort: they convey no sense of success or failure.

Lest the reader be misled by the presentation: the above breakdown would also occur in a subgoal-goal structuring of the actors' plans using formally defined **IN-ORDER-TO** links (see above and Schmidt,1978): Relevant actions are put in immediate **IN-ORDER-TO** relation to the major goals, while irrelevant actions are either not related at all or are at least one action removed from the action **IN-ORDER-TO** achieve the major goal. Since plan recognition structure is a dynamically changing knowledge representation, multiple final achievements are possible if we look at the perceptions on a moment by moment basis.

The Belief Hypothesis. The belief hypothesis has to do with the perceived clarity of the action: "How clear is it that the actor *intended* to do that particular action?". A number of studies have confirmed the importance of the observer forming such belief structures for plan recognition and story understanding. In contrast to the plan hypothesis which can be appreciated from an analysis of verbal expressions as well as physical events, it is difficult (and I think, impossible) to appreciate the belief hypothesis from an analysis of verbal expressions, since natural English is so laden with supposition. It is easier to appreciate this hypothesis if one visualizes assigning evaluations to physical events -- which is, of course, what this research is all about. Again I identified three categories of action in decreasing order of probability:

1. **Unambiguous**. The actor clearly wanted to ACHIEVE the state or first-order change specified in the CONceptualization. Hitting actions and door manipulations (door closing, opening) are usually instances of such actions.

2. **Ambiguous**. Many actions, such as most simple traveling actions, convey ambiguity or no clear sense that the actor wanted to ACHIEVE a particular state (did the actor want to leave the place or go to another?). Such actions are sometimes seen in retrospect: we see some movement, but its significance (viz., the *action*) does not become apparent until later.

3. **Unrelated.** Some few actions were unrelated to any actor's goals: For example, leaving the screen cannot be a goal of an actor, since the actor does not know where the screen is.

As with the previous hypothesis, this one is directly related to a rigorous formulation (see Bruce and Newman (2)). Also like the previous hypothesis, we found that this classification of the actions against the film is a natural one.

Results. For each hypothesis, a multivariate regression analysis was used to predict the probability of a point of definition for the one second intervals over the 71 second film. The use of this analysis is quite straightforward. As in any regression analysis, the values of the dependent variable are predicted by an equation which weights the values of one or more independent variables. The regression analysis will derive an optimal measure of correlation, or best-fit, between the values of the independent variables and the values of dependent variable. In our study two regression analyses, are compared against each other to see which best fit accounts for when people see actions. The analyses have the same dependent variable but different independent variables.

The dependent variable for both analyses is the observed probability of a point of an action definition. This takes on a value for each one second interval over the film. Newtson obtained the 71 dependent values on that variable by summing the button presses by all 55 respondents within each interval. These sums are shown in Figure 1. This distribution is typical of distributions obtained in other Newtson's studies. It is obvious that points of action definition are well agreed on by the 55 respondents and that the distribution of points of action definition is highly variable and therefore interesting. A single person only rarely indicates two points of action definition in an interval, so this dependent variable very closely approximates the probability that a person will report an action in a one second interval. It is the function of the regression analyses to estimate how well the hypotheses will predict when and with what probability a person will report seeing an action.

The independent variables for each of the two regressions was a count of the number of points of action definition provided by the machine reports for each corresponding one second interval. The machine can provide many such points of definition in a single second, whereas a person is physically limited (by reaction times and perhaps decision processes). Therefore the regression was done as a quadratic approximation to correct for asymptotc, non-linear, response characteristics. The quadratic polynomial approach was planned before the data was seen. By using a quadratic model (which squares the independent variable) we *allow* that this may mean the person presses only once despite rapid perceptions, although the model will still reveal the linear component since it contains the nonsquared version of the independent variable as well.

The Plan Hypothesis and the Belief Hypothesis represent distinct models and are evaluated in separate regressions. Each hypothesis classifies actions reported by the computer program into three mutually exclusive categories for each one second interval. Since each category has a squared and a nonsquared version, this provides six independent (linearly weighted) variables to predict the values of the dependent variables.

The regressions were done in a stepwise fashion to give an indication of which variables are independently predictive of observed points of action definition. The order of entry of the variables with cumulative variance accounted for in the dependent variable are:

- *Plan Hypothesis*:
    1. LT-Succeed (18%)
    2. ST&SC-Succeed (21%)
    3. ST&SC-Succeed^2 (24%)
    4. LT-Succeed^2 (25%)
    5. Irrelevant^2 (26%)
    6. Irrelevant (26%)
- *Belief Hypothesis*:
    1. Unambiguous (33%)
    2. Unambiguous^2 (53%)
    3. Ambiguous (54%)
    4. Ambiguous^2 (54%)

Figure I-2: Frequency Distribution for Points of Action Definition over the Heider and Simmel Film.

5. Unrelated (54%)[21]

These results are clear. Both hypotheses are successful in predicting the distribution of points of action definition shown in Figure 1, but the belief hypothesis accounts for about 100% more of the variation than the plan hypothesis. In fact, we can collapse the success in prediction to a single factor. The unambiguous action factor of the belief hypothesis alone accounts for most variance (53%).

An examination of where the plan recognition hypothesis failed and the belief hypothesis succeeded suggested that the failure occurred primarily because physical-object-related actions like opening and closing a door (but not moving from one place to another)despite their merely instrumental roles in the actors plans had high probabilities of eliciting a point of definition. This characteristic alone seems to have made the difference in the two situations.

These results have an impact on the design of an expectation-based system for action perception. The success of the belief hypothesis suggests that expectation rules should work with knowledge of the actors and objects in the situation. This knowledge should permit a simple-minded assessment of intentional actions. It may be possible, in many applications, to formulate simplified rule systems which rely on heuristics and logical constraints to adjust the expectation set. The heuristics would focus less on the plans and ultimate goals of the actors than on locally relevant actions -- what goals an actor is likely to have in a physical situation. Thus, in a room with a door, the system may posit door closing and opening routinely regardless of hypotheses about actors' plans. Perhaps the visual system should be set up to detect obviously intentional actions without regard to current hypotheses about what an actor is trying to accomplish. This is analogous to a lexical lookup as that process is employed Schank's (1975) request-based parsing. In certain situations, it may be possible to use the objects in view to index action schemas. The observations suggest that people are more simple minded in the act of perceiving than they might be in reflecting on what they have perceived. Perhaps plan recognition is more a phenomenon of *explaining* actions, as Wilensky (1978) suggests, than for preparing for them. Such results could be taken as a caveat not to try to make artificial systems for action perception too smart, since the human existence proof is not yet established for mechanically ideal plan recognition in visual perception.

The NEXT component was constructed in the expectation that the plan hypothesis would succeed and the belief hypothesis would fail. This predicts a pattern of low probability actions followed by a higher probability action through the joint probabilities of two action patterns which reflect different perceptual organizations: (a) schemata for all the actions expressed individually, and (b) schemata which link the actions by the NEXT component but thereby register only the point of action definition at the last action in the sequence. The conceptualizations expressed by the two patterns would be different but related by the **IN-ORDER-TO** relationship. For example, the set of actions which can be phrased "the small triangle moved to the door, he opened the door, and then the small circle escaped" would have the alternative plan organization roughly phrased "the small triangle moved to the door in order to open it and release the small circle." However, with the relative failure of the plan recognition hypothesis, we cannot yet rely on the NEXT component as a mechanism for structuring the perception process. In the computer analysis of the film this component was relegated to the less compelling role of handling curving trajectories for

[21]Not enough instances to warrant squaring for quadratic.

actors, as in the action "the small triangle and circle went around the box". The use here was dictated more by the choice of first-order change descriptors than theoretical motivations.

Future Research

Cartoon animation systems.

Cartoon animation systems permit an animator to create simple animations in two or three dimensional space for computer analysis and for real-time play-back. Such systems have been developed for AI studies of cartoon generation (Kahn , 1977, 1979). Of more importance than graphic realism (which seems to be the current trend) is that the system create long films which tell visually compelling stories.

The major advantage of a computer system for playing back the film is that the physical events can be investigated in depth using action schemas or, perhaps, more complete versions of the system for artificial perception outlined in this paper. Another major advantage is that cartoon animations are already digitilized, whereas other schemes require an extensive, and expensive, commitment of human labor.

Natural environment systems

There can be no question that cartoon animations impose a severe handicap on the quality of the study of action perception. Similar difficulties are apparent in other domains of computer vision and understanding. Animations provide action illusions, just like drawings provide object illusions. Newtson's work, in contrast to the present work, deals with records of natural events (except for the one study with the Heider and Simmel film (Massad, Hubbard and Newtson, 1979). State descriptions and first-order change descriptions for human and animal motions can borrow from various movement notation schemes (Newtson, 1976; Newtson and Engquist, 1976). These characterize continuous movements (first-order changes) as rotations about joints and the like. However, scoring films for such movements is extremely tedious and time consuming. When matters turn to computing first-order changes in natural environments automatically, the definitions of these changes become much more complex than the definitions permitted with cartoon animations.

Ultimately, an action recognition machine should be used on natural imagery, and not be dependent on cartoons. The problem is that this approach appears to demand that we wait until work in object and motion perception is completed to provide us with the machines which will interpret the objects and motions. In the animation system this is not a problem because the films can be completely encoded, objects pre-labeled, and motions defined using simple schemes. The closest we are likely to come to mundane situations where experience in action perception for natural human actions can be gained appears to be surveillance situations. The required characteristics are (a) fixed (unobtrusive, fixed perspective) camera position and (b) well-defined and limited domains of application.

The "image processor" for computing state and first-order change information for surveillance systems also suggests how useful but incomplete object and motion information can be obtained in natural environments. This processor capitalizes on fixed camera position by permitting hand-segmentation and hand conceptual-labeling of the background scene along with manual input of perspective criteria. The state description processor computes segmented occlusions of the background as possible un-labeled objects. The positions of the un-labeled objects can be described relative to the positions of known, conceptually labeled, objects (as in, "an object is near the door"). Movement and direction of occlusion boundaries could provide much of the first-order change information. Movement detected within the occluding object by time-down gradient operators may provide for further segmentation. I envisage that the output of this image processor is a partial description similar to the one generated for our experimental study.

It is easy see how to fool such a processor, so it is important to select surveillance situations carefully. Where there is informed deceptive action the surveillance devices would need to be "less smart", at least originally (see the discussion on "primacy effects"), but some forms of safety-related, unobtrusive, or deception-hardened surveillance remain interesting. Such contexts motivate further exploration of the expectation mechanism in its role in shifting and focusing attention for verbal description. Of course, for better or worse, such talk of exploiting real world circumstances leads to talk of "HAL 9000" computers (from the movie "2001: A Space Odyssey" c. 1969) that keep watch on peoples' activities.

Language and Perception Mappings

The relationships between language and perception have been under discussion for many years. A most famous hypothesis about this relationship is the Whorfian Hypothesis (Whorf, 1956);[22] that *language provides the structure to perception*. In present terms, the CON's available for action schemas are constrained to be CON's derived for simple English (not necessarily Russian or Hopi), and their presuppositions and entailments constrain the COLLECT and ACHIEVE propositions. We have already seen how logico-linguistic analysis of English expressions relates to action schema derivations. More experience is required with the present paradigm to properly evaluate the status of the Whorfian Hypothesis within it.

One idea for developing a system for artificial perception of actions would be to develop a brute force Whorfian perceptual machine. This has an extreme advantage over some alternatives: The mapping to English is planned into the design of action recognition schema. It is wise to anticipate the problem of English description because the mapping is so complex (see Waltz, 1980) , for additional insights). An added advantage is, perhaps, that the burden on what the system should see can be placed on the users of the system. The users describe, in English, what the system should see, and the system is then engineered to see it. This tactic has worked well in the implementation of practical systems for natural language understanding.

The serious fallacy in such a Whorfian perceptual machine is that presuppositions and entailments of English expressions may over-specify action schemas. This was the case in the protocol study. In a two-dimensional world where surfaces are not hidden from view, this is surprising. In a three-dimensional world view, inferences would have to carry part of the load in the computation. The expectation mechanism is one reasonable place to adjust over-specified action schemas against the realities of perspective viewing.

Object and Motion Perception

In this study I have de-emphasized work in object and motion perception (and their kin) in order to emphasize action perception. I noted that *interpreting* objects and motions may not be required in interpreting actions. The general idea is that environmental cues may be used in perceiving actions that permit hypotheses about the appropriate conceptualizations for objects and motions.

---

[22]Also called the Sapir-Whorf Hypothesis (Leech, 1974), linguistic determinism, and other things.

Object and motion conceptualizations could be elicited not only by bottom up cues but by recognized actions as well. This idea is very much in the spirit of recent work on motion perception which seeks to show how object recognition can be facilitated by motion computations (Ullman, 1979). The action schema mechanism can, in principle, resolve ambiguities about objects and motions (after actions are recognized), but it remains to be seen whether this avenue is worth exploring.

A more serious concern arises in the assumptions made about the computations of state and first-order change descriptions. Basically, I have assumed that these computations are carried out independently of the computations for action schemata. The only defense for this assumption is that it is often reasonable, and (b) it certainly makes it easier to gain experience with the physical antecedents of action perception. But then, the present approach is clearly deficient in dealing with the perception of actions in static forms (Birdwhistell, 1970; Herman, 1980) .

This paper began with the thesis that actions are perceived "directly", yet the *action postulate* asserts that actions are defined in second-order changes. The issue is whether the account of how actions are perceived need reference to how motions and objects are perceived. In the present formulation, motions are derivative in first-order change descriptions but they are not the same as first-order change descriptions. For an action to be derivative in second-order change, which is then derivative in first-order change does not imply that we have studied motions. We have not developed a theory of selection in first-order change, only one for second order change. If this strategy is fruitful, then the meaning of "direct perception of actions" is preserved. Action, motion, and object perception are different phenomena in terms of their computation.

A Derivative Scheme for Robotics

An interesting research domain is the artificial perception of actions by mobile robots. This domain would permit one to exploit the derivative or indirect aspect of action perception. The necessity of a fixed camera position is relaxed because the robot navigation system can provide object knowledge and perspective knowledge to the image processor. Unfortunately, it is much less practical to hand segment the scene for its stable (unchanging) parts as proposed for surveillance systems, since this would have to be done from many perspectives.

However, robot navigation schemes are now being designed that navigate the robot through the use of internal world models. In effect, the robot vision (and sensing) system maintains an internal model of the external world. Having designed the artificial perception of actions to work with derivative, state and first-order change information allows us to consider action perception from the internal world models. This strategy should facilitate the implementation of action perception in robots.

2. Acknowledgement

I had direct knowledge of Newtson's work over several years, and this paper represents a brief, nine-month, opportunity I had to work directly with him and some of the data he had collected over the years. I would like to thank Darren Newtson and John Rotondo for an enjoyable few months collaborating on a topic we all find intriguing. The work has been done, on the sly, with internal funding from The University of Virginia and the Carnegie-Mellon University Robotics Institute.

References

II. Appendix: State and First-Order Change Language

This describe the state and first-order change finite language composed for the Heider and Simmel (6) film. All numeric information is implicit in the symbolic descriptors and all assignments are deterministic. The language is not arbitrary and reflects heavily on the protocol analysis and the frame-by-frame analysis of the geometric and perceptual idiosyncrasies of the film. It is therefore particular to the film. It is included here for completeness.

The objects, O, in the film are a small circle (SC), small triangle (ST), large triangle (LT), a box (BX) and a door into it (DR). The variable P stands for any of the persons, SC, ST, or LT. The variable I stands for any of the inanimate objects, BX or DR. The variable R stands for any region as seen in Figure I-1. Figure I-1 also shows labeled containment boundaries (defined below), the unit of measurement, and the actors to scale. The following propositions are defined:

(Touch P O) An irreflexive proposition recorded when any line segment of the actor, P, physically touches any line segment of the object, O.

(Break I) An object, I, is broken when the line segments of I are broken.

(Rotate P) ST or LT rotated in either direction to any degree.

(Line-of-sight P P) An irreflexive proposition recorded when a line drawn between the centroids of the two P's does not intersect any line segment of any I. (This seems perceptually correct for the film.)

(Door Open/Closed/Opening/Closing)
Closed requires that the outer end of the door be touching the box; the definitions of Open, Opening, and Closing are transparent.

(At P P) Symmetric and Irreflexive: A actor, P, is at the location of another P if they are within 7 units of each other's centroids and there is a line-of-sight between them.

(At P R) A actor, P, is at the location, R, if his centroid lies in an area identified in Figure I-1, where R is offscreen (OF), outside the box (OT), inside the box (IN), the wall (WL), the doorway (DW), the door (DR), a lower-left region (LL), a bottom region (BT), a lower-right region (LR), a back region (BK), an upper-right region (UR), and a top region (TP). Note, by implication, if an actor is both IN and LR this places him in the lower-right corner inside the box.

(Near P P) A symmetric and irreflexive relation between two actors if their centroids are within 14 units and there is a line-of-sight between them and they are not at each others location.

(Move P Slow/Normal/Fast)
Translation of centroid by thresholding at 1.5 and 4.0 units.

(Moveto,MovePast,MoveFrom P P)
For two actors, P, there is an asymmetric, irreflexive relation: if there is a line of sight one actor will move toward, past, or from the other. The method is to determine the vector angle between the direction of movement and the direction of the other actor with the reference in the moment before the present moment (law of cosines). The angles are in 120^{o} arcs.

(Moveto,MoveFrom P R)
For an actor and a location, R, there is movement toward and from but not past. Movement possibilities are provided as labeled containment boundaries (double arrows) in Figure 1: offscreen-to-left (OFLF), offscreen-to-top (OFTP), offscreen-to-bottom (OFBT), offscreen-to-back (OFBK), outside (OT), inside (IN), wall (WL), door (DR), lower-left (LL), lower-right (LR), upper-right (UR), top (TP), back (BK), and bottom (BT). The method determines the nearest labeled containment boundary intersected by the directed movement vector of the actor. As in directional movement

between actors, where an actor was in the last frame determines where he is seen to be moving to or from in the present frame. The moveto vector is computed as if he was at the last frame location and the movefrom vector is computed as if he is at the present frame location. To illustrate, if an actor is above the box and moves rightward his movement is interpreted as going to the BK, not OFBK or UR. If he moves left, his movement is interpreted as going OFLF.

Figure II-1: Film Frame Worksheet for Appendix I.

Table of Contents

List of Figures