Reasoning about the world as perceived by an agent

Erik Sandewall *
Department of Computer and Information Science
Linköping University
Linköping, Sweden
E-mail: ejs@ida.liu.se

Abstract

The relationship between quantitative/physical knowledge on one hand, and qualitative/conceptual knowledge on the other hand is discussed, in particular from the viewpoint of the design of autonomous agents. It is proposed to view quantitative knowledge as the primary model of the world, and qualitative knowledge as a secondary world model. The transformation from primary to secondary model is seen as a kind of perception transformation, and may be described as an automaton. It is proposed that an equational style of specification, called discontinuity equations, is appropriate for expressing such transformations. Results regarding criteria for the existence of a unique solution to discontinuity equations, are briefly reviewed.

1 Conceptual and physical knowledge

1.1 Conceptual vs. physical knowledge

An intelligent autonomous agent that moves and acts in the real world must be able to deal with both qualitative and quantitative information about its environment (and about itself). The qualitative knowledge, by conventional wisdom, should use a conceptual structure that is reminiscent of what we find ourselves using, for example in natural language. It should use concepts and constructs such as "objects", "actions", "events", temporally dependent "properties" of objects, and so forth. The formal character of the conceptual structures may be as wff in a suitable logic, or as partial interpretations for logic formulas, or simply as high-level data structures according to taste.

It is equally clear that the quantitative knowledge must be organized according to the principles of classical engineering, with mathematics (especially calculus), physics, and automatic control engineering in successive layers providing both concepts and theory.

*This research was supported by the Swedish Board of Technical Development and the IT4 research program

It is not at all clear, however, how the qualitative, concept-oriented knowledge is *related* to the quantitative, physics-oriented knowledge, and what is the best kind of formal description of such a relationship. The character of that relationship is very significant both from a practical viewpoint (when designing intelligent robotic systems), and theoretically in order to have a precise definition of all aspects of intelligent agents. The purpose of the present paper is to address that issue and propose a well founded approach.

1.2 Time continuous fluents

We choose to describe all aspects of behavior as *time-continuous fluents*, i.e. functions of time where time is measured as real (or possibly rational) numbers. The reason for this choice is that continuous time is necessary in order to accomodate differential equations and other aspects of mathematical calculus, which are the standard tool for expressing quantitative knowledge. For compatibility we assume that qualitative knowledge also refers to continuous time. A qualitative fluent such as "John is having dinner" is therefore seen as a function from the real time axis to truth-values. Since the range of such qualitative fluents is a finite and discrete set, qualitative fluents must be piecewise constant. In other words they are time-continuous but amplitude-discrete.

Our topic can then be rephrased as follows: what is the relationship between the amplitude-continuous, quantitative fluents and the amplitude-discrete, qualitative fluents which occur in the quantitative resp. the qualitative accounts of one and the same scenario in the real world?

1.3 Temporal span in qualitative fluents

It is important to notice that qualitative fluents which capture conceptual information, can *not* be seen simply as predicates on the current value of a corresponding quantitative fluent. For example, the conceptual fluent

"John is having dinner at the current time" can not be determined by only taking a snapshot of the state of the world itself at time t; one must also take the state of the world during a preceding interval into account. John might for example be interrupting his dinner for a brief phone call around time t, so that the snapshot catches him doing something that is unrelated to the dinner.

For another example, consider the conceptual fluents "car A is overtaking car B" and "car A is driving alongside with car B". Clearly each of these propositions would qualify as a fluent, so that in a temporal logic using explicit time one should be able to say for example "during the interval between time 56 and time 82, car A is overtaking car B", presumably expressed with a formula such as

$$[56, 82]Overtakes(A, B)$$

The quantitative fluents for the automobiles can also be easily defined, and would include for each of the cars its position on the road as a function of time. However again it would not be possible to define the transformation from quantitative to qualitative fluent on a momentary basis, since a snapshot of the two cars at one point in time is not sufficient for determining whether car A is overtaking or driving alongside car B at that time.

1.4 Proposal: automata as characterizers

In view of the observations in the previous sections, the following viewpoint is now proposed:

- A distinction is made between a *primary* and a *secondary* model of the world in itself
- The primary world model is quantitative, as described by physical or other scientific knowledge. For example in the car scenario, the primary world model would include quantitative fluents specifying the position and other parameters of the cars as functions of time.

  In general, each momentary world state $s(t)$, in the primary world model, is constructed from objects having a number of fluents such as mass, velocity, acceleration, temperature, and so on. Physical knowledge characterizes the primary world state using the familiar concepts of calculus and physics.
- The secondary world model contains qualitative/conceptual knowledge, and characterizes the world state and its history *indirectly* and *in aggregated form*. It is obtained from the primary world model using a transformation with "memory" which I shall call a *characterizer*, and which can be informally thought of as a kind of perception function.

  In the most general formulation, the secondary world model at time t can be seen as $p(S(t))$, where the historical world state $S(t)$ is defined as

  $$S(t) = \{u \mapsto s(u) | u \leq t\}$$

  i.e. as the history of the world up to the present point in time. More specifically, however, the historical information is captured by the agent's current state, which is dependent not only on its current sensory input but also on its past history.

The reason for informally viewing the secondary world model as the result of perception is as follows. Consider a primitive device (or animal) which has the capabilities of perception, action, and a limited amount of computation for going from perception to action. It would be reasonable to design (or evolve) such a device as a real-time finite-state automaton, whose actions at each point in time is determined by its current, discrete internal state as well as by the quantitative value of its current inputs. The design would therefore contain one component we could speculatively call a perceiver, which maps continuous sensory input to a piecewise constant fluent for internal state, and another component we could call an effector which maps (continuous sensory input) $\times$ (discrete internal state) to (signals to actuators). – We ignore at this time the sensory limitations, and approximate $s(t)$ with the sensory input at time t.

When intelligence is added, in a next step of device evolution, it is natural to let the intelligence *modify* rather than replace the low level behavior that was just described. Using a temporal logic or by other means, the intelligence should *reason about* the intervals during which the piecewise constant fluents in the internal state have their distinct possible values. In particular this gives the device a capability for temporal prediction, which serves for making plans for sequences of actions that exploit opportunities, or avoid accidents or other disadvantages in its environment.

We use the term "characterizer" rather than "perceiver" for the transformation from continuous sensory fluents to discrete conceptual fluents, since the analogy with perception is only heuristic, and the design as such has a mathematical and intended engineering character.

If time were seen as discrete, then the transformation from the primary to the secondary world model could therefore be written as an equation

$$C(t) = \phi(S(t), C(t-1))$$

as illustrated in figure 1. The internal state $C(t)$ that is obtained from the characterizer sub-system is then used both to control the choice of action in the lower behavioral level of the layered agent, and as the basis for reasoning and other cognitive activities. The transformation

that is performed in the characterizer uses a current state which is updated in each step. Below we will revise this formulation so that it works for continuous time.

The internal state of the intelligent device is then seen as a real-time knowledge base, and conceptual structures are thought of as those discrete structures which are useful to have in the internal state of an autonomous agent. One consequence of this view is that to understand why qualitative knowledge has the structure it has, it is important to understand the natural designs of characterizers. Another consequence is that formal description methods for characterizers are significant for A.I.

1.5 Situated automata

The view proposed above is related to but distinct from the view of situated automata proposed by Rosenschein [Ros85, RK86], according to which it is possible to let the internal state remain unanalyzed. Rosenschein defines knowledge in an agent as follows: the agent is said to *know* a proposition φ iff φ is true in all external states $S(t)$ such that $p(S(t))$ equals the current internal state of the agent, where p is a function characterizing (but not necessarily implementing) the agent. The idea is that the internal state of the agent, or situated automaton, should be seen as an implementation issue.

In situated automata, *both* physical and conceptual knowledge therefore refer to the world state. There is no connection between the evaluation of a proposition φ in $S(t)$, and the actual computation of the internal state $p(S(t))$. In my proposal, on the other hand, the only way to evaluate the proposition φ in $S(t)$ is to evaluate it in the agent's or automaton's internal state $C(t) = p(S(t))$.

2 Formal specification of characterizers

If conceptual information is seen as the result of "perception" by characterizers, then it is important to find concise modes of expressing and analyzing such characterizers. Two obvious choices of formal approach come to mind: *direct use of logic*, and the use of *automata approaches*. In the former alternative, one uses a declarative formulation of the relationship between the historical state $S(t)$ of the environment, and its qualitative description. In the latter approach one defines an automaton which can operate over continuous time, and whose state $C(t)$ contains the qualitative information for time t.

For a very simplistic example, the propositional fluent "having dinner", applied to a person, would be characterized by an automaton which takes observations of the world as input, and where one component in the internal state is $\mathcal{T}$ during those intervals where the observed person is having dinner.

2.1 Case study

Consider the simplest possible case of a characterizer with internal state, namely the one illustrated in figure 2. The input signal is a real number, and the output signal is a truth-value, both of course as functions of time. The output classifies the input as "high" or "low", using the two thresholds a and b where $a < b$ to determine when the input "becomes" low and "becomes" high, respectively.

In a conventional logic formulation, the criterium would be expressed as follows:

$$d(t) \geq b \Rightarrow h(t)$$

$$d(t) \leq a \Rightarrow \neg h(t)$$

$$\exists t_0[d(t_0) = a \wedge Inside(t_0, t)] \Rightarrow \neg h(t)$$

$$\exists t_0[d(t_0) = b \wedge Inside(t_0, t)] \Rightarrow h(t)$$

where the auxiliary predicate *Inside* is defined by

$$Inside(t_0, t) \Leftrightarrow t_0 < t \wedge \forall t'[t_0 < t' \leq t \Rightarrow a < d(t') < b]$$

This expression is actually still not complete, but suffices as the basis for discussion. In the other approach one characterizes the transformation from d to h as in figure 1, where there is an automaton (the characterizer) with a feedback loop, so that the old value of the h fluent is used for defining the new value. The characterizer ϕ is specified by

$$h_o = [d \leq a \mapsto \mathcal{F} \mid d \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto h_i]$$

where h_o is the output fluent from the characterizer, and d and h_i are the input fluents. The operator [...] is defined like in LISP, so that $x = [\alpha \mapsto y \mid \vartheta]$ is equivalent to $(\alpha \Rightarrow x = y) \wedge (\neg\alpha \Rightarrow x = [\vartheta])$. The feedback loop means that h_o and h_i satisfy

$$h_i(t) = h_o(t - 1)$$

which can be rewritten more compactly as $h_i = \Lambda h_o$, where the previous value operator Λ is defined by $\Lambda f(t) = f(t - 1)$. The total transformation from d to $h = h_o$ is therefore the (or a) solution of the equation

$$h = [d \leq a \mapsto \mathcal{F} \mid d \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto \Lambda h]$$

combined with a requirement on the initial value $h(0)$.

The effect of the third branch of the conditional is that if the "previous" value of h is known then the "next" value is also known; the effect of the first two branches is to specify when and how the value is to change.

2.2 Discussion

We have now seen two descriptions of the transformation from the quantitative world description, which in this case is the continuous-valued fluent d, to the qualitative world description which in this toy example is the truth-valued fluent h. In comparison, the first approach uses logic in order to refer directly to earlier time, i.e. to the full range of information in $S(t)$, and is clearly not a viable alternative due to its complexity. The second approach essentially views the characterizer as a real-time automaton where the fluent h_o serves both as the internal state and as the output, and in addition the behavior of the automaton is characterized equationally.

The automaton viewpoint is the obvious one to use, due both to its conciseness and because it corresponds to the intuitions – the state of the automaton can be identified with the basis for the secondary, conceptual world model as discussed above.

There are also other ways of defining automata, besides the equational style. In the present context one might define the characterizer by a set of assignment rules, such as "when the input signal becomes $\geq b$, set the output to $\mathcal{T}$". Alternatively one might define it in two steps, where (for the example) the first step is a "memory-less" classification of the input into the three alternatives $d \geq b$, $d \leq a$, and $a < d < b$, and the second step is a 3x2 transition table.

The pros and cons of those alternatives are discussed more extensively in the full version of the present paper, which is available from the author. Briefly, however, we are interested in an equational description because of its formal similarity with differential equations, which are a primary means of describing continuous physical processes. For those cases where discrete and continuous change is tied together, one would expect to obtain more concise descriptions if both kinds of change can be accomodated in the same, equational framework.

There is however one significant problem which still has to be resolved: the characterizer function that was defined above relies on the discreteness of the time domain, since the operator Λ refers to the value at the previous time-point. The approach should therefore be modified to consider the continuous-time equivalent of Λ which refers to the "previous" value over an infinitesimal time-step. This is the topic of the next section.

3 Discontinuity equations

3.1 Characterizers

The generalization to continuous time is straight-forward, and relies on the following basic assumptions. We consider fluents which are functions from real time, to some continuous domain (for example real numbers, or vectors thereof), and which are *piecewise continuous*. In other words, for each fluent h there is a set of real numbers, called breakpoints, such that $h(t)$ is continuous for all t which are not breakpoints, and every finite interval contains only a finite number of breakpoints.

The continuous counterparts of Λ are defined as follows. If f is a fluent, then λf is another fluent such that

$$\lambda f(t) = \lim_{t' \to t-0} f(t')$$

which in particular means that outside breakpoints, $\lambda f(t) = f(t)$. For example, if f is piecewise constant, has a breakpoint for $t = a$, and is b_1 for $t < a$ and b_2 for $t \geq a$, then λf is b_1 for $t \leq a$ and b_2 for $t > a$. The "infinitesimal delay" affects the value in the breakpoint a itself, but not outside the breakpoint.

The left limit operator λ is matched by a corresponding right limit operator ρ defined by

$$\rho f(t) = \lim_{t' \to t+0} f(t')$$

Furthermore we assume for each fluent f that besides piecewise continuity, $\lambda f(t)$ and $\rho f(t)$ shall be defined for every t.

An equation with one or more fluents as unknowns and which uses the operators λ and ρ will be referred to as a *discontinuity equation*. It makes sense to use λ as the infinitesimal counterpart of Λ, and by direct generalization one would expect the discrete-time formulation above to generalize to

$$h = [d \leq a \mapsto \mathcal{F} \mid d \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto \lambda h]$$

Below we introduce a standard form and criteria for discontinuity equations which guarantee the existence and uniqueness of the solution. In this example the standard form would be

$$h = \rho h = [d \leq a \mapsto \mathcal{F} \mid d \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto \lambda h]$$

However before turning to the formal part let us consider also another example of a transformation which can be characterized using discontinuity equations, and now a task which uses a simple "data structure".

3.2 Example: continuous set accumulation

Consider the scenario described in figure 3. Fluents may have "car id" (for example the registration plate number) and "set of car ids" as values. A number of observer automata c_1, c_2,... take suitable input, and have as output the car id of the car that is being observed or was

most recently observed by the automaton. Each c_i is assumed to be similar to the output fluent in the previous example, satisfying $c_i = \rho c_i$. The output of the stack of accumulation automata should be a set-valued fluent whose value at each point in time is the set of all car ids that have been or are being observed by any of the observer automata.

If each accumulator $f_o = \phi(c, f_i)$ is defined as

$$\phi(c, f_i) = \{c\} \cup f_i$$

then the whole set accumulation process is the solution of the discontinuity equation

$$f_n = \rho f_n = \lambda f_n \cup \bigcup_i c_i$$

which conforms to the same normal form as was mentioned in the previous example.

4 Determinacy of discontinuity equations

From a theoretical point of view we would first of all be interested in knowing whether a discontinuity equation has exactly one solution for every choice of the dependent fluent (such as d in the first example above) and of the initial value for the solution. In such a case the discontinuity equation should be called *deterministic*, by analogy with the corresponding automata. Determinism is only obtained, as we shall soon see, if some additional requirements are imposed.

We would have a two-way interest in knowing the criteria for determinacy: in some cases determinism is desired and then we want to know what are the requirements for having it; in other applications one may wish to represent non-determinism and one is interested in the complementary requirements.

In this conference paper we only give a summary and simplified account of the determinacy results. The complete proofs are available in a separate paper[San90].

Let $\mathcal{D}_i$ be a continuous or discrete domain, and $\mathcal{D}_o$ be a discrete domain. A *fluent* f is a piecewise continuous mapping from the positive real numbers to $\mathcal{D}_i$ or to $\mathcal{D}_o$, for which λf and ρf exist at all times. A function

$$\phi : \mathcal{D}_i \times \mathcal{D}_o \to \mathcal{D}_o$$

is called a *characterizer from $\mathcal{D}_i$ to $\mathcal{D}_o$*. Characterizers will be used as the right-hand side of discontinuity equations on standard form. If ϕ is a characterizer from $\mathcal{D}_i$ to $\mathcal{D}_o$ and $v \in \mathcal{D}_o$, then the *eigendomain* of ϕ for v is defined as

$$\{x \mid \phi(x, v) = v\}$$

i.e. the set of those $x \in \mathcal{D}_i$ for which ϕ maps v to itself.

A characterizer ϕ is said to be *regular* iff the following two conditions are satisfied:

1. For every v and x,

$$\phi(x, \phi(x, v)) = \phi(x, v)$$

2. The union of the eigendomains for all v in $\mathcal{D}_o$ equals $\mathcal{D}_i$.

The first condition is equivalent to saying that x is in the eigendomain of $\phi(x, v)$, for all x and v. A characterizer ϕ is said to be *deterministic* iff it is regular, and in addition one of the following conditions applies:

1. For every $v \in \mathcal{D}_o$, its eigendomain is an open set i.e. does not include its edge
2. $\mathcal{D}_i$ is a discrete set

A discontinuity equation is in *standard form* iff it has the form

$$h = \rho h = \phi(d, \lambda h)$$
$$h(0) = \phi(d(0), h(0))$$

where ϕ is a regular characterizer, d is a known fluent where $d = \rho d$, and h is the unknown fluent. The first line of the equation is taken as an abbreviation for

$$\forall t[0 \leq t \Rightarrow h(t) = \rho h(t)] \ \wedge$$
$$\forall t[0 < t \Rightarrow h(t) = \phi(d(t), \lambda h(t))]$$

and it is easily seen that for every combination of fluents d and h it is well defined whether this condition holds or not.

We have proved[San90] the following

Lemma 1 *Every fluent h which is a solution of a discontinuity equation on standard form satisfies*

$$h = \phi(d, h)$$

Theorem 2 *The standard form discontinuity equation with deterministic characterizer ϕ has at least one solution, and exactly one solution h for every choice of $h(0)$.*

In the example above, one would define ϕ as follows:

$$\phi(x, \mathcal{T}) = (x > a)$$
$$\phi(x, \mathcal{F}) = \neg(x < b)$$

where the relations $<$ and $>$ are seen as truth-valued functions. Thus the eigendomain of ϕ for $\mathcal{T}$ in this example is $\{x \mid x > a\}$, and the eigendomain for $\mathcal{F}$ is $\{x \mid x < b\}$. Therefore the eigendomain of ϕ for $\mathcal{T}$ consists of those x such that if the input fluent is x, the output fluent retains the value $\mathcal{T}$ if it already has it. Likewise, the eigendomain for $\mathcal{F}$ consists of those x for which the output fluent is able to retain the value $\mathcal{F}$. The

two eigendomains are overlapping, open sets, and their union covers the whole positive axis.

We can then re-obtain the tentative formulation of the discontinuity equation for the first example, previously stated in section 3.1 above. The expressions for ϕ above can be equivalently re-written as

$$\phi(x, \mathcal{T}) = [x \leq a \mapsto \mathcal{F} \mid \mathcal{T} \mapsto \mathcal{T}]$$

$$\phi(x, \mathcal{F}) = [x \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto \mathcal{F}]$$

and again as the deterministic characterizer

$$\phi(x, v) = [x \leq a \mapsto \mathcal{F} \mid x \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto v]$$

so that the discontinuity equation $h = \rho h = \phi(d, \lambda h)$ becomes in this example

$$h = \rho h = [d \leq a \mapsto \mathcal{F} \mid d \geq b \mapsto \mathcal{T} \mid \mathcal{T} \mapsto \lambda h]$$

which is what we had in section 3.1.

In theorem 2 both the requirement $h = \rho h$ and the condition on $h(0)$ are essential, as the following examples show. Let $\phi(x, v)$ be as above, and consider the input fluent d defined by $d(t) = t$. The only corresponding solution h is

$$h(t) = [t < b \mapsto \mathcal{F} \mid \mathcal{T} \mapsto \mathcal{T}]$$

but if the condition $h = \rho h$ is dropped then for example the fluent

$$h(t) = [t \leq (a + b)/2 \mapsto \mathcal{F} \mid \mathcal{T} \mapsto \mathcal{T}]$$

also satisfies the remaining equation. Also consider the input fluent d defined by $d(t) = a + t$. The only corresponding solution h is

$$h(t) = [t < b - a \mapsto \mathcal{F} \mid \mathcal{T} \mapsto \mathcal{T}]$$

since a is only in the eigendomain for $\mathcal{F}$, not for $\mathcal{T}$. However the fluent which is constantly $\mathcal{T}$ does satisfy $h = \rho h$ and $h = \phi(d, \lambda h)$, and is only rejected if the condition on $h(0)$ is included.

5 Conclusions

The present version of the paper is very brief due to the page constraints for the conference proceedings. A more extensive version of the paper is available from the author.

John Hallam's comments about an earlier draft are gratefully acknowledged.

References

[RK86] Stanley J. Rosenschein and Leslie Pack Kaebling. The synthesis of digital machines with provable epistemic properties. In *Proceedings of the Workshop on Theoretical Aspects of Reasoning about Knowledge, Monterey, CA*, pages 83–98, 1986.

[Ros85] Stanley J. Rosenschein. Formal theories of knowledge in ai and robotics. *New Generation Computing*, 3:345–357, 1985.

[San90] Erik Sandewall. Transformations on fluents, and discontinuity equations. Technical Report LAIC-IDA-90-TR7, IDA, 1990.

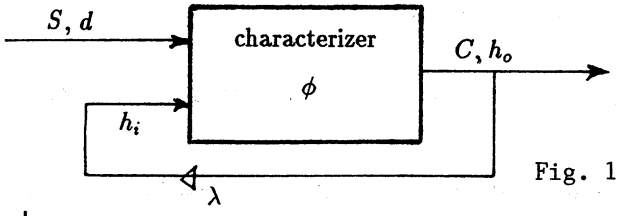

Fig. 1

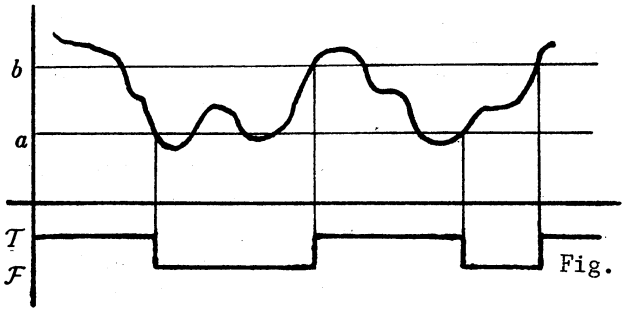

Fig. 2

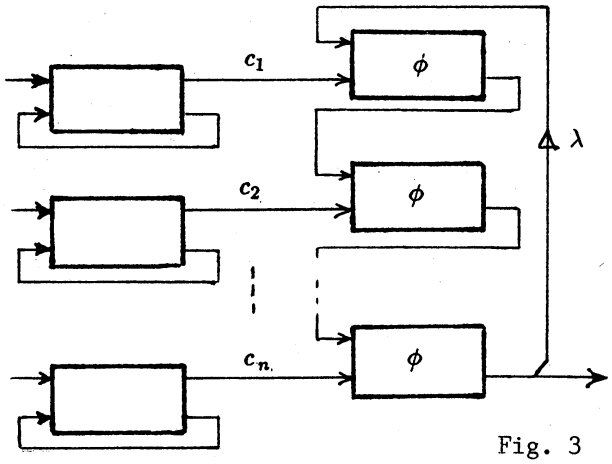

Fig. 3