Towards Diagnosing Hybrid Systems

Sheila McIlraith, Gautam Biswas
Knowledge Systems Lab
Stanford University
Stanford, CA 94305
{sam,biswas}@ksl.stanford.edu

Dan Clancy, Vineet Gupta
Caelum Research Corporation
NASA Ames Research Center
Moffett Field, CA 94035
{clancy,vgupta}@arc.nasa.gov

Abstract

This paper reports on the findings of an on-going project to investigate techniques to diagnose complex dynamical systems that are modeled as hybrid systems. In particular, we examine continuous systems with embedded supervisory controllers which experience abrupt, partial or full failure of component devices. The problem we address is: given a hybrid model of system behavior, a history of executed controller actions, and a history of observations, including an observation of behavior that is aberrant relative to the model of expected behavior, determine what fault occurred to have caused the aberrant behavior. Determining a diagnosis can be cast as a search problem to find the most likely model for the data. Unfortunately, the search space is extremely large. To reduce search space size and to identify an initial set of candidate diagnoses, we propose to exploit techniques originally applied to qualitative diagnosis of continuous systems. We refine these diagnoses using parameter estimation and model fitting techniques. As a motivating case study, we have examined the problem of diagnosing NASA's Sprint AERCam, a small spherical robotic camera unit with 12 thrusters that enable both linear and rotational motion.

1 Introduction

The objective of our project has been to investigate how to diagnose hybrid systems – complex dynamical systems whose behavior is modeled as a hybrid system. Hybrid models comprise both discrete and continuous behavior. They are typically represented as a sequence of piecewise continuous behaviors interleaved with discrete transitions (e.g., (Branicky 1995)). Each period of continuous behavior represents a so-called *mode* of the system. For example, in the case of NASA's Sprint AERCam, modes might include *translate_X-axis*, *rotate_X-axis*, *translate_Y-axis*, etc. (Alenius & Gupta 1998). In the case of an Airbus fly-by-wire system, modes might include *take-off*, *landing*, *climbing*, and *cruise* (Sweet 1995). Mode transitions generally result in changes to the model governing the continuous behavior of the system, as well as to the state vector that initializes that behavior in the new mode. Discrete transitions that dictate mode switching are modeled by finite state automata, temporal logics, switching functions, or some other transition system, while continuous behavior within a mode is modeled by ordinary differential equations (ODEs) or differential and algebraic equations (DAEs).

While at the macroscopic level, all physical systems are inherently continuous, the exploitation of hybrid models, and in particular the distinguishing of modes and discrete mode transitions proves useful for modeling and analysis of many physical systems. For example, discrete supervisory controllers embedded in continuous systems may impose multiple continuous modes of operation that are best modeled as hybrid systems. Hybrid models are also useful for simplifying models of complex system behavior. Many complex systems exhibit fast nonlinear behaviors that are hard to model and analyze. A number of these fast transients can be attributed to parasitic parameters, whose values are hard to estimate. In such cases, nonlinear system behavior can be abstracted to piecewise continuous behaviors with discrete transitions that are simpler to analyze and interpret (e.g., (Mosterman & Biswas 1997a)). In the examples above, hybrid models use discrete transitions to model both *controller actions*, and so-called *autonomous jumps*, i.e. model-induced jumps from one continuous behavior to another (Branicky 1995). As we shall see in this paper, we may also use discrete transitions to model *exogenous actions*, i.e., unpredicted actions that cause components of our system to fail.

The problem we address in this paper is how to diagnose such hybrid systems. For the purposes of this paper, we consider the class of hybrid systems that are continuous systems with an embedded supervisory controller, but whose hybrid models contain no autonomous jumps. The class of systems we consider can be modeled as a composition of a set of component subsystems, each of which is itself a hybrid system. We assume that the system operation is being tracked by a monitoring and observer system (e.g., (Mosterman & Biswas 1999a)) that ensures that the system behavior predicted by the model does not deviate significantly from the observed behavior in normal system operation. When observations occur outside this range, the behav-

ior is deemed to be aberrant and diagnosis is initiated. In this paper, we consider faults whose onset is abrupt, and which result in partial or complete degradation of component behavior. The general problem we wish to address can be stated as follows: *Given a hybrid model of system behavior, a history of executed controller actions, a history of observations, including observations of aberrant behavior relative to the model, isolate the fault that is the cause for the aberrant behavior*. Diagnosis is done online in conjunction with the continued operation of the system. Hence, we divide our diagnosis task into two stages, initial conjecturing of candidate diagnosis and subsequent refinement and tracking to select the most likely diagnoses.

In this paper we conceive the diagnosis problem as a *model selection, fitting, and comparison problem.* The task is to find a mathematical model and associated parameter values that best fit the system data. These models further dictate the components of the system that have malfunctioned, their mode of failure, the estimated time of failure and any additional parameters that further characterize the failure. To address this diagnosis problem, we propose to exploit AI techniques for qualitative diagnosis of continuous systems to generate an initial set of qualitative candidate diagnoses and associated models, thus drastically reducing and simplifying the size of the model search space. This is followed by parameter estimation and model fitting techniques to select the most likely mode and system parameters for candidate models of system behavior, given both past and subsequent observations of system behavior and controller actions. The main contributions of the paper are:

- formulation of the hybrid diagnosis problem;
- the exploitation of techniques for qualitative diagnosis of continuous systems to reduce the diagnosis search space; and
- the use of parameter estimation and data fitting techniques for evaluation and comparison of candidate diagnoses.

In Section 2 we present a formal characterization of the class of hybrid systems we study and the diagnosis problem they present. This is followed in Section 3 by a brief description of NASA's Sprint AERCam, which we have used as a motivating example and which we will use to illustrate certain concepts in this paper. In Section 4 we describe the algorithms we use to achieve hybrid diagnosis. The generation of initial candidate qualitative diagnoses is described in Section 4.1, and the subsequent quantitative fitting and tracking of candidate diagnoses and their models is described in Section 4.2. Finally in Section 5, we summarize and discuss where our investigation will go from here.

2 Problem Formulation

In this section we provide a formal definition of the class of hybrid systems we study in this paper, and define the hybrid diagnosis problem.

Definition 1 (Hybrid System) *A hybrid system is a 5-tuple $\langle \mathcal{M}, X, \mathcal{F}, \Sigma, \phi \rangle$, where*

- *$\mathcal{M}$ is a finite set of modes $(\mu_1, \ldots, \mu_k)$, representing the possible modes of system behavior.*
- *$X \subseteq R^n$ defines the continuous state variables. $x(t)$ describes the continuous behavior at time t.*
- *$\mathcal{F}$ is a finite set of functions $\{f_{\mu_1}, \ldots, f_{\mu_k}\}$, such that for each mode, μ_i, $f_{\mu_i}(t, x(t)) : R \times X \rightarrow X$ defines the continuous behavior of the system in μ_i.*
- *Σ is a finite set of discrete actions $(\sigma_1, \ldots, \sigma_l)$, which transition the system from one mode to another.*
- *ϕ is a transition function which maps an action, mode and system state vector into a new mode and initial state vector, i.e., $\phi : \Sigma \times \mathcal{M} \times X \rightarrow \mathcal{M} \times X$.*

Definition 2 (System State) *The state of a hybrid system at time t is defined by the discrete mode and the continuous state at that time, as represented by the tuple $(\mu_i^t, x(t))$.*

To define the hybrid diagnosis problem, we augment the description of our hybrid systems as follows.

Definition 3 (Hybrid System Diagnosis Terminology) *Consider a hybrid system $\langle \mathcal{M}, X, \mathcal{F}, \Sigma, \phi \rangle$ comprised of m potentially malfunctioning components $(c_1, \ldots, c_m)$, each of which is itself a hybrid system.*

- *$\mathcal{M}_\mathcal{F} \subseteq \mathcal{M}$ is a distinguished subset of modes representing fault modes of the hybrid system. There is at least one fault mode $\mu_j \in \mathcal{M}_\mathcal{F}$, for each component c_i. For notational convenience, we will use the notation μ_F to denote a fault mode $\mu_j \in \mathcal{M}_\mathcal{F}$.*
- *We assume that transitions to fault modes are achieved by exogenous actions. Hence, Σ, the finite set of discrete actions is divided into two subsets such that $\Sigma = \Sigma_c \cup \Sigma_e$, and*
  - *Σ_c is a finite set of controller actions, and*
  - *Σ_e is a finite set of exogenous actions.*

  We define a controller action history, $\mathcal{A}$ to be a sequence of time-indexed actions performed by the controller.
- *$X_{obs} \subseteq X$, denotes the continuous state variables that are observable. $x_{obs}(t)$ denotes the values of observations at time t. We define the observation history, O to be values of $x_{obs}(t)$ at a sequence of sample times, t_i.*
- *For each continuous behavior function of a fault mode $f_{\mu_F} \in \mathcal{F}$, we distinguish parameters θ_F of the function, which are to be estimated as part of the diagnosis task. Allowable ranges may be associated with some or all of the individual parameters. These parameters will, e.g., characterize the degree of degradation of some aspect of component behavior.*

- *a Model, Mod, for time-indexed mode sequence $[\mu_1, \ldots, \mu_m]$ is the corresponding time-indexed piecewise continuous sequence of functions $[f_{\mu_1}, \ldots, f_{\mu_m}]$.*

In this paper we make several simplifying assumptions regarding our diagnosis task. In particular, we make a single-time fault assumption. We assume that our systems do not experience multiple sequential faults. Further, we assume that faults are abrupt, resulting in partial or full degradation of component behavior. We also assume that components fail independently. This is of course, not always a reasonable assumption.

Intuitively, we can think of our hybrid diagnosis task as a big model-finding, model-fitting and model-comparison problem. The behavior of the system as it transitions through controller-induced and fault-induced modes μ_i can be modeled by the appropriate sequence of functions, f_{μ_i}. Hence, given infinite resources, we could, in principle, build a sequence of functions, corresponding to a model for every possible sequences of modes, and estimate parameters to maximally fit the observed data to each model. The model with the best fit would indicate the state and mode history of the system, including any fault modes that had occurred. Clearly this is not a computationally feasible approach, particularly since fault modes can occur at potentially infinitely varying times and with many different parameter values.

Instead, we propose to monitor observed system behavior against one model, Mod_{normal}, the model for the mode sequence $[\mu_1, \ldots, \mu_m]$ that corresponds to the mode sequence dictated by the controller action history $\mathcal{A}$, the initial state $x(t_0)$, and the transition function ϕ. We define the probability that the system is operating according to the normal or expected model, given the action and observation history, $P(Mod_{normal} \mid \mathcal{A}, \mathcal{O})$, as the measure of fit of the observation history $\mathcal{O}$ with the model.

When aberrant behavior is detected, e.g., when observations fall outside a range predicted by the Mod_{normal}, we assume that the normal model does not reflect the evolution of system behavior, and the diagnosis task commences. Given a hybrid system $\langle \mathcal{M}, X, \mathcal{F}, \Sigma, \phi \rangle$, a controller action history, $\mathcal{A}$ and a history of observations, $\mathcal{O}$ which includes observations of aberrant behavior, the **hybrid diagnosis task** is to determine what components are faulty, what fault mode caused the aberrant behavior, when it occurred, and what the values of the parameters associated with the fault mode are. In the AERCam system, a diagnosis might be that thruster T_1 experienced a blockage fault at time t_i, and that the thruster is operating at 50% its normal level.

Again, we are faced with an enormous search problem to determine the time-indexed sequence of parameterized functions that best fits the observed data. To overcome this challenge, this paper proposes the exploitation of qualitative reasoning techniques to prune the search space. In particular, from the controller action history $\mathcal{A}$, we initially assume the system is operating normally, as dictated by the model Mod_{normal}, with associated mode history $[\mu_1, \ldots, \mu_m]$, temporally indexed with the corresponding controller action times from $\mathcal{A}$. Exploiting previous research on temporal causal graphs for qualitative diagnosis of continuous systems (Mosterman & Biswas 1999b), we compute a set of candidate qualitative diagnoses that are consistent with the model and associated mode history $[\mu_1, \ldots, \mu_m]$ and the observed aberrant behavior. More formally,

Definition 4 (D-tuple) *A D-tuple is a 4-tuple $\langle C, \mu_F, t_F, \theta_F \rangle$, where μ_F is a fault mode, t_F is the time the fault mode commenced, θ_F is the parameter values associated with the fault mode behavior, and C is the set of failed components corresponding to μ_F.*

Definition 5 (Candidate Qualitative Diagnosis) *Given a hybrid system $\langle \mathcal{M}, X, \mathcal{F}, \Sigma, \phi \rangle$, an action history $\mathcal{A}$, a model and associated mode sequence $[\mu_1, \ldots, \mu_m]$, and a history of observations, $\mathcal{O}$ which includes observations of aberrant behavior, D-tuple $\langle C, \mu_F, t_F, \theta_F \rangle$ is a candidate qualitative diagnosis iff there exists a range of parameter values $\theta_F = [\theta_l, \theta_u]$, and time range $t_F = [t_l, t_u]$ such that the occurrence of fault mode μ_F with parameter values θ_F in time range t_F is consistent with $\mathcal{O}$ and $\mathcal{A}$.*

Hence, a candidate qualitative diagnosis stipulates a fault mode, corresponding to one or more faulty components. It also stipulates a lower and upper bound, $[t_l, t_u]$, on the time the fault mode occurred. This range generally corresponds to the start times of the controller induced modes preceding and following the fault, or up to the point the fault was detected. For example, if we are conjecturing that the fault occurred within mode μ_i of the mode sequence associated with the normal model, then t_F would be the interval $[t_i, t_{i+1}]$, where t_i and t_{i+1} are the start times for modes μ_i and μ_{i+1}, respectively. The parameter range, $[\theta_l, \theta_u]$ is generally the reals, unless otherwise constrained. An algorithm for computing candidate qualitative diagnoses in discussed in Section 4.1.

Each candidate qualitative diagnosis, also indirectly dictates a new candidate mode sequence and a new candidate model – $[\mu_1, \ldots, \mu_i, \mu_F, \mu_{F_{i+1}}, \ldots, \mu_{F_m}]$ and $[f_{\mu_1}, \ldots, f_{\mu_i}, f_{\mu_F}, f_{\mu_{F_{i+1}}}, \ldots, f_{\mu_{F_m}}]$, respectively. The new candidate mode sequence corresponds to the previous mode sequence $[\mu_1, \ldots, \mu_m]$ with the fault mode μ_F interjected at t_F. Note that the occurrence of fault mode μ_F may affect subsequent modes in the mode sequence, as dictated by transition function ϕ. Hence all modes in the new candidate mode sequence that follow μ_F reflect the modes obtained from the controller actions of $\mathcal{A}$ transitioning from this faulty mode and continuous state.

Observe that since each candidate qualitative diagnosis only conjectured ranges for the time of the fault

mode, t_F and parameter values associated with the fault mode, θ_F, the associated candidate models are underconstrained. To complete our model, we must find a more precise estimate of the time of failure, t_F, and the parameter values, θ_F. In Section 4.2, we discuss two methods for estimating t_F and θ_F. The first uses expectation maximization (EM) (e.g., (Dempster, Laird, & Rubin 1977)) to estimate θ_F and t_F simultaneously. The second uses statistical hypothesis testing methods (e.g., (Basseville & Nikiforov 1993)) to estimate t_F, and then applies nonlinear regression methods to estimate θ_F. Both proposed techniques have advantages and disadvantages. We are currently exploring the efficacy of these techniques in practice.

Each candidate qualitative diagnosis $\langle C, \mu_F, t_F, \theta_F \rangle$ has now been refined to refer to an individual time point t_F and parameters θ_F, and we have associated a unique candidate model, which we refer to as Mod_C, with each diagnosis. From this, we define a candidate diagnosis to be a D-tuple and associated candidate model Mod_C that has a posterior probability, given the observations and controller actions, that is above some specified threshold value. In particular,

Definition 6 (Candidate Diagnosis) *Given a hybrid system $\langle \mathcal{M}, X, \mathcal{F}, \Sigma, \phi \rangle$, a history of controller actions $\mathcal{A}$, and a history of observations $\mathcal{O}$, D-tuple $\langle C, \mu_F, t_F, \theta_F \rangle$ with associated model Mod_C is a candidate diagnosis for the hybrid system, iff Mod_C is consistent with $\mathcal{A}$ and $P(Mod_C \mid \mathcal{O}) > \alpha$, for defined threshold value $\alpha \in [0, 1]$.*

Bayes Theorem provides us with the mathematics to estimate both the posterior probability of the parameters, given the observation history $\mathcal{O}$ and the model Mod_C, i.e.

$$P(\theta \mid \mathcal{O}, Mod_C) = \frac{P(\theta \mid Mod_C) P(\mathcal{O} \mid \theta, Mod_C)}{P(\mathcal{O} \mid Mod_C)},$$

where the normalizing constant $P(\mathcal{O} \mid Mod_C)$ is defined as

$$\begin{aligned} P(\mathcal{O} \mid Mod_C) &= \int P(\mathcal{O}, \theta \mid Mod_C) d\theta \\ &= \int P(\mathcal{O} \mid \theta, Mod_C) P(\theta \mid Mod_C) d\theta. \end{aligned}$$

Bayes Theorem also provides us with the mathematics to estimate the posterior probability of the model given the observation history, i.e.,

$$P(Mod_C \mid \mathcal{O}) = \frac{P(\mathcal{O} \mid Mod_C) P(Mod_C)}{P(\mathcal{O})},$$

where $P(\mathcal{O} \mid Mod_C)$ is the measure of fit of $\mathcal{O}$ to Mod_C, $P(Mod_C)$ is the prior $P(\mu_F)$, and $P(\mathcal{O})$ is a normalizing constant.

The observation history $\mathcal{O}$ contains a history of time indexed observations $[x_{obs}(0), \ldots, x_{obs}(t - \delta)]$, which we use to initially estimate parameters and to compute the posterior for the candidate models. As the system progresses, we obtain further observations, and we update the probabilities of our candidate models for every subsequent observation $x_{obs}(t)$, exploiting a Markov assumption.

$$P(Mod_C \mid x_{obs}(t)) = \frac{P(x_{obs}(t) \mid Mod_C) P(Mod_C)}{P(x_{obs}(t))},$$

where $P(x_{obs}(t) \mid Mod_C)$ is the measure of fit of $x_{obs}(t)$ to Mod_C, $P(Mod_C)$ is, by a Markov assumption, $P(Mod_C \mid \mathcal{O})$ computed as the result of the previous observations, $[x_{obs}(0), \ldots, x_{obs}(t - \delta)]$, and $P(x_{obs}(t))$ is again a normalizing constant.

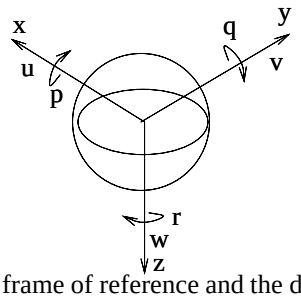
Intuitively, candidate diagnoses are D-tuples whose associated models characterize the observed system behavior within some threshold of accuracy. In order to compare candidate diagnoses, we must compare their associated candidate models. It is insufficient to simply choose the model with the best fit because such a criterion is likely to be biased in favor of a more complex, highly parameterized, model which can overfit the data. To compare different models, we use Bayesian model comparison as described in (MacKay 1991). As noted by MacKay, Bayesian model comparison captures the notion of Occam's Razor, favoring simpler models over more complex models. In terms of diagnosis, Bayesian model comparison captures the commonly held bias in model-based diagnosis of preferring minimal diagnoses, i.e., diagnoses with the minimal number of failing components (e.g., single fault hypotheses).

3 Motivating Example: The AERCam

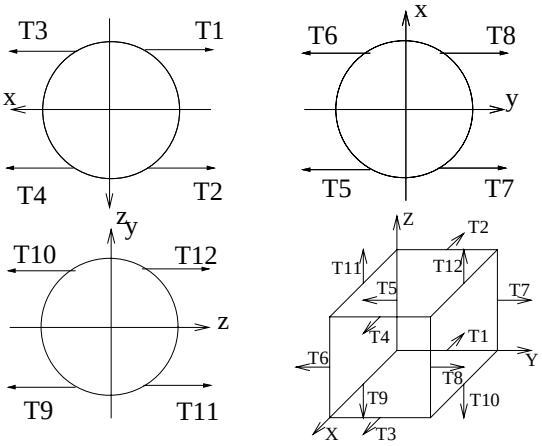
We are using NASA's Sprint AERCam and a simulation of system dynamics and the controller written in HCC (Alenius & Gupta 1998) as a testbed for investigating monitoring and diagnosis techniques in hybrid environments. We describe the dynamic model of the AERCam system briefly, a more detailed description of the models appears in (Alenius & Gupta 1998).

The AERCam is a small spherical robotic camera unit, with 12 thrusters that allow both linear and rotational motion (Figure 1). For the purposes of this model, we assume the sphere is uniform, and the fuel that powers the movement is in the center of the sphere. The fuel depletes as the thrusters fire.

The dynamics of the AERCam are described in the AERCam body frame of reference. The translation velocity of this frame with respect to the shuttle inertial frame of reference is 0. However, its orientation is the same as the orientation of the AERCam, thus its orientation with respect to the shuttle reference frame changes as the AERCam rotates (i.e., it is not an inertial frame). The twelve thrusters are aligned so that there are four along each major axis in the AERCam body frame (Figure 1). For modeling purposes, we assume the positions of the thrusters are on the centers of the edges of a cube circumscribing the AERCam. Thrusters T_1, T_2, T_3, T_4 are parallel to the X-axis and are used for translation along the X-axis or rotation around the Y-axis. Firing thrusters T_1 and T_2 results in translation along the positive X-axis, and firing thrusters T_1 and



The body frame of reference and the directions of velocities (u, v, w) are the components of the translation velocity. (p, q, r) are components of the angular velocity.



Three views of the AERCam, showing the thrusters, and showing all the thrusters together in the cube circumscribing the AERCam.

Figure 1: The AERCam axes and thrusters

T_4 to get a negative rotation around the Y-axis. Similarly, thrusters T_5, T_6, T_7, T_8 are parallel to the Y-axis, and are used to rotate around the Z-axis, and thrusters $T_9, T_{10}, T_{11}, T_{12}$ are parallel to the Z-axis, and are used for rotation around the X-axis. AERCam operations are simplified by making it either translate or rotate. Thrusters are either on or off, therefore, the control actions are discrete. In normal mode of operation, only two thrusters are on at any time. For safety of the crew and the shuttle equipment, the thruster linear and angular velocities are not allowed to exceed prespecified thresholds.

3.1 AERCam dynamics

A simplified model of the AERCam dynamics based on Newtonian laws is derived using an inertial frame of reference fixed to the space shuttle (Etkin & Reid 1995). The AERCam position in this frame is defined as the triple (x, y, z). Let $\vec{\mathbf{V}}$ be the velocity in the AERCam body frame, with its vector components given by (u, v, w). The frame rotates with respect to the inertial reference frame with velocity $\omega = (p, q, r)$, the angular velocity of the AERCam. The rotating Body frame implies an additional Coriolis force acting upon the AERCam. We assume uniform rotational velocity since in the normal model of operation, the AERCam does not translate and rotate at the same time (Arnold 1978, pg. 130). Similar equations can be derived for the rotational dynamics (Alenius & Gupta 1998).

$$d(m\ \vec{\mathbf{V}})/dt = \vec{\mathbf{F}} - 2m(\vec{\mathbf{V}} \times \vec{\omega}) \qquad \text{Newton's Law}$$
$$\vec{\mathbf{V}}\ dm/dt + md(\vec{\mathbf{V}})/dt = \vec{\mathbf{F}} - 2m(\vec{\omega} \times \vec{\mathbf{V}})$$

The resultant equation for each coordinate appears below.

$$du/dt = F_x/m - 2(qw - vr) - (u/m) * dm/dt$$
$$dv/dt = F_y/m - 2(ru - pw) - (v/m) * dm/dt$$
$$dw/dt = F_z/m - 2(pv - qu) - (w/m) * dm/dt$$

3.2 Position Control Mode of the AERCam

In the position control mode, the AERCam is directed to go to a specified position and point the camera in a particular direction. Assume the AERCam is at position A and directed to go to position B. In the first phase, the AERCam rotates to get one set of thrusters pointed towards B. These are then fired, and the AERCam cruises towards B. Upon reaching close to B, it fires thrusters to converge to B, and then rotates to point the camera in the desired direction.

To facilitate the illustration of the diagnosis problem, we use a simple trapezoidal controller, which we explain in two dimensions. Suppose the task is to travel along the x-axis for some distance, then along the y-axis. Such a manoeuvre could be needed in the space shuttle, to avoid hitting some objects. In order to do this, the AERCam fires its x thrusters for some time. Upon reaching the desired velocity, these are switched off. When the AERCam has reached close tox the desired point, the reverse thrusters are switched on, and it is brought to a halt — the velocity graph is a trapezium. The same process is repeated in order to travel along the y direction.

4 Diagnosing Hybrid Systems

In Section 2 we defined the restricted class of hybrid systems we wish to diagnose, and the hybrid diagnosis problem. In this section we discuss one method for computing hybrid diagnoses. In particular, in this paper we propose to exploit previous work on qualitative diagnosis of continuous systems to help diagnose hybrid systems. The benefit of qualitative techniques in this context is that they use qualitative representations of the domain knowledge to drastically reduce the search space for candidate diagnoses and hence candidate models. In Section 4.1 we discuss a technique for generating candidate qualitative diagnoses, and their associated candidate models of system behavior, first proposed for qualitative diagnosis of continuous systems (Mosterman & Biswas 1999b). In Section 4.2 we discuss techniques for model fitting and for model (and hence diagnosis) comparison. In particular we discuss

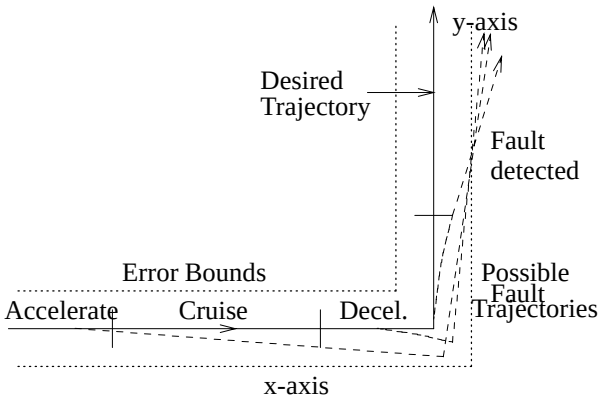


Figure 2: Trajectories of AERCam under various possible faults. The fault trajectories are simplified for illustration purposes.

techniques for estimating the parameters of the candidate models, and the likelihood of the models, and for continued monitoring and refinement of the candidate models as the system continues to operate and observations continue to be made.

We illustrate these techniques with the following simple AERCam example. Consider the scenario depicted in Figure 2. In the first acceleration phase, the AERCam is being powered by thrusters $T1$ and $T2$. Assume that at some point in this phase, a sudden leak in the $T2$ thruster causes an abrupt change in its output. As a consequence, the AERCam starts veering to the right of the desired trajectory, as illustrated by the left-most dotted lines in Figure 2. (The other dotted lines represent other potential candidate diagnoses consistent with the point of detection of the failure.) Soon after this occurs, the supervisory controller commands the AERCam to turn off Thrusters $T1$ and $T2$ with the objective of getting the AERCam to cruise in a straight line. In the faulty situation, the AERCam has some residual angular velocity about the z-axis, so it continues to rotate in the cruise mode. Then the controller turns on thrusters $T3$ and $T4$, to decelerate the AERCam with the objective of bringing it to a halt. Again, this objective is not entirely achieved in the the faulty situation. Next, thrusters $T5$ and $T6$ are switched on, to move the AERCam in the y direction. However, since the AERCam is not in the desired orientation after the failure, the position error due to faulty thruster $T2$ accumulates causing a greater and greater deviation from the desired trajectory of the system. The position of the AERCam is being continuously sensed, filtered for noise and monitored. At some point within the y translation the trajectory crosses a predefined error bound and is flagged by the monitoring system as aberrant behavior relative to model Mod_{normal}. At this point, the diagnosis task begins.

4.1 Qualitative Candidate Generation

Given the normal system model Mod_{normal}, a history of controller actions $\mathcal{A}$ and associated mode sequence $[\mu_1, \ldots, \mu_m]$, and a history of observations $\mathcal{O}$ including one or more observations of aberrant behavior, we wish to generate a set of consistent *candidate qualitative diagnoses* $\langle C, \mu_F, t_F, \theta_F \rangle$, and *associated models* as described in Definition 5. To do so, we extend techniques for generating qualitative diagnoses of continuous dynamic systems to deal with hybrid systems with multiple modes. A full description of the model representation and propagation mechanism applied to continuous systems diagnosis can be found in (Mosterman & Biswas 1997b; 1999b).

In the case of our AERCam example, the action history $\mathcal{A}$ is [(on($T1$), on($T2$)), (off($T1$), off($T2$)), (on($T3$), on($T4$)), (off($T3$), off($T4$), on($T5$), on($T6$)), (off($T5$), off($T6$))]; the mode sequence is [$accelerate\_x$, $cruise\_x$, $decelerate\_x$, $accelerate\_y$, $cruise\_y$]. Mod_{normal} is the time-indexed sequence of functions [$f_{accelerate\_x}$, f_{cruise_x}, $f_{decelerate\_x}$, $f_{accelerate\_y}$, f_{cruise_y}] derived from the system dynamics overviewed in Section 3. The time indexing corresponds to the times of the control actions. Finally, the observation history $\mathcal{O}$ is sequence of $(x(t_i), y(t_i), z(t_i))$ and computed velocity and acceleration at the sample times t_i.

To generate candidate qualitative diagnoses we construct an abstract model of the dynamic system behavior, Mod_{normal} as a temporal causal graph. A part of the temporal causal graph for the AERCam dynamics is shown in Figure 3. The graph expresses directed cause-effect relations between component parameters and the system state variables. Links between variables are labeled as: (i) $+1$, implying direct proportionality, (ii) -1, implying inverse proportionality, and (iii) $\int$, implying an integrating relation. An integrating relation introduces a temporal delay in that a change on the cause side of the relation affects the derivative of the variable on the effect side. This adds temporal characteristics to the relations between variables. Some edges are labeled by variables, implying the sign of the variable in the particular situation defines the nature of the relationship.

The candidate generation algorithm is invoked for every initial instance of an aberrant observation. The aberrant observation plus the controller action history $\mathcal{A}$ are input to a backward propagation algorithm that operates on the temporal causal graph. The algorithm operates backward in time from μ_m, the last mode in the given mode sequence $[\mu_1, \ldots, \mu_m]$:

1. For the current mode, extract the corresponding temporal causal graph model, and apply the *Identify Possible Faults* algorithm. Details of this algorithm are presented in (Mosterman & Biswas 1999b), but the key aspect of this algorithm is to propagate the aberrant observation expressed as a $\pm$ value, backward depth-first through the graph. For example, given that the $y-$position of the AERCam has deviated $-$

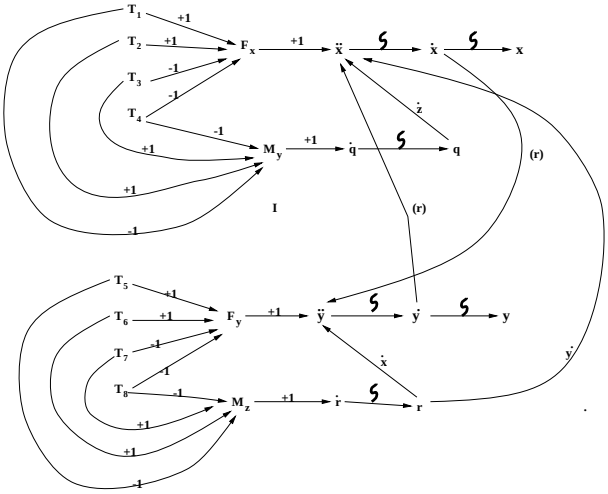


Figure 3: A subset of the temporal causal graph showing the relations between Thrusters $T1 - T8$ and the x and y positions of the AERCam.

(i.e., below normal), backward propagation implies $d(y)/dt$ is $-$, and so on, till we get T_5^- and T_6^-, implying thrusters $T5$ and $T6$ are possible faulty with decreased thrust performance. Propagation along a path can terminate if conflicting assignments are made to a node. The goal is to systematically propagate observed discrepancies backward to identify all possible candidate hypotheses that are consistent with the observations. In our example, the component parameters, $T1 - T12$ form the space of candidate hypotheses.

2. Repeat Step 1 for every mode in the mode sequence, to μ_1. The system model needs to be substituted as the algorithm traverses the mode sequence backwards, therefore, back propagation will be performed on a different temporal causal graph for each mode in the controller history[1].

The output of this step is a set of qualitative diagnoses $\langle C, \mu_F, t_F, \theta_F \rangle$, each with an associated candidate mode sequence and candidate model, as described in Section 2. Returning to our AERCam example, three qualitative candidate diagnoses are generated[2]. The first candidate diagnosis is that $T2$ failed in the x acceleration phase, and that there was a jump to a new mode called $bad\_T2\_accelerate\_x$. The time of the fault mode transition is $[t_1, t_2]$, and the parameters associated with the failure – the percentage degradation of the component is in the range $[0, 100]$. So the first candidate qualitative diagnosis is $\langle T2, bad\_T2\_accelerate\_x, [t_1, t_2], [0, 100] \rangle$. The candidate mode sequence is $[accelerate\_x, bad\_T2\_accelerate\_x, cruise\_x, decelerate\_x, accelerate\_y, cruise\_y]$, and the associated candidate model is defined accordingly. The second candidate qualitative diagnosis is that $T4$ failed in the deceleration phase of x translation, i.e., $\langle T4, bad\_T4\_decelerate\_x, [t_3, t_4], [0, 100] \rangle$. The third candidate is that $T6$ failed during y acceleration, i.e., $\langle T6, bad\_T6\_accelerate\_y, [t_4, t], [0, 100] \rangle$, where $t = t_D$, the time of detection of the aberrant behavior.

[1] Heuristics may be introduced to cut off back-propagation along the mode sequence beyond a time limit. The rationale would be that any significant fault manifestation unless masked, would produce observable changes in the state variables within this pre-specified time limit.

[2] Based on the assumption that the thrusters do not fail positively, i.e., their output cannot exceed their 100% maximum thrust value as defined by parameter restrictions in the model.

4.2 Model Fitting and Comparison

The candidate qualitative diagnoses and the associated candidate mode sequences and candidate models provide a qualitative characterization of the hypothesized faults, obtained through a qualitative analysis of the model of normal behavior, Mod_{normal}, and the observations. Given this information, the next phase of the diagnosis process is quantitative refinement of the qualitative candidate diagnoses and their associated models through parameter estimation and data fitting, followed by tracking of the fit of subsequent observations to the candidate models. The goal is to identify a unique diagnosis, or barring that, to provide a probabilistic ranking of the plausible candidates, so that the supervisory controller can use this information in making decisions on future action selection.

As observed in the previous section, the model associated with the qualitative candidate diagnosis, Mod_C is underconstrained. Both the time of the fault mode occurrence, t_F and the parameters associated with the faulty behavior θ_F are represented as ranges and must be estimated. Further, the candidate qualitative diagnoses were generated from initial observations of aberrant behavior, and their consistency can be further evaluated by monitoring the qualitative transients associated with each candidate. The refinement process is performed by a set of *trackers* (Rinner & Kuipers 1999), one for each candidate diagnosis and associated model. Each tracker comprises both a *qualitative transient analysis* component and a quantitative *model estimation*, component as shown in Figure 4. The two components operate in parallel as described below.

Qualitative Transient Analysis

The qualitative transient analysis component performs a further qualitative analysis of the consistency of candidate qualitative diagnoses based on monitoring of higher-order transients whose manifestation is seen over a longer period of time. If the transients of a candidate qualitative diagnosis do not remain consistent with subsequent observations, the candidate diagnosis will be eliminated and the *model estimation* component informed. The technique we employ is derived from techniques for qualitative monitoring of continuous systems

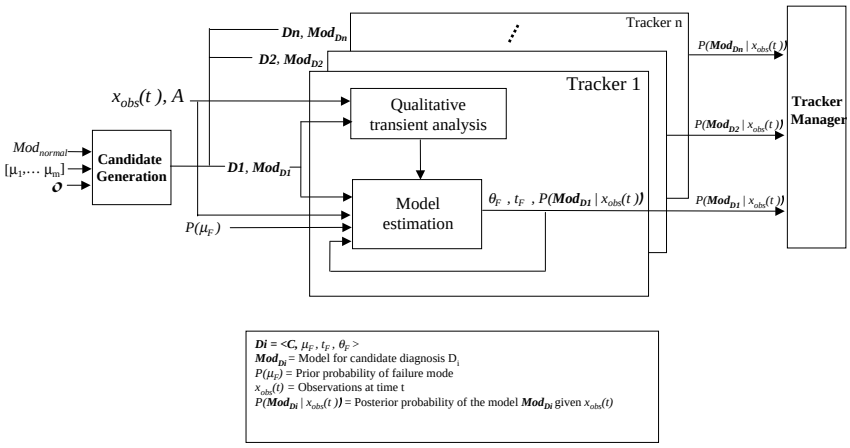


Figure 4: Candidate generation, refinement and tracking

as described in (Mosterman & Biswas 1997b; 1999b; Manders, Mosterman, & Biswas 1999).

Given a candidate qualitative diagnosis, $\langle C, \mu_F, t_F, \theta_F \rangle$, the temporal causal graph and causal propagation machinery described in Section 4.1 are used to compute the qualitative dynamic, transient behavior for all the observed variables. Predicted future behavior is expressed as a qualitative *signature* of magnitude (i.e., instantaneous (0^{th} order)), slope (i.e., 1^{st} order time derivatives), and higher-order effects. Details of the *Predict Future Behavior* algorithm appear in (Mosterman & Biswas 1999b). In brief, the algorithm forward propagates the effect of a hypothesized fault along the temporal causal graph in a breadth-first manner to build the fault signature for individual observations. For example, the predicted signature for a $T6$ failure, i.e., $T6^-$, for observed position y is $\langle 0, 0, -1, -1(d(x)/dt) \rangle$, and for x is $\langle 0, 0, 0, -1(r) \rangle$. A $T6$ failure introduces a negative second order deviation in the y position value, but a third-order positive deviation in the x position value (because the fault causes a negative r value). A mode change governed by a controller action may cause the signs of $d(x)/dt$ or r to change, which would then reverse the third order effect of the $T6$ failure on y, and x. This information is used to evaluate the consistency of the candidate diagnoses with respect to the transient characteristics of subsequent observations, and to reject inconsistent candidate qualitative diagnoses. In such a case, the corresponding tracker is eliminated, and the remaining candidate probabilities are normalized accordingly.

Model Estimation

The purpose of the model estimation component is to perform quantitative model fitting, i.e., to provide a quantitative estimate of the parameters of the models and to assign a probability to each of the candidate models (and hence candidate diagnoses), given the noisy observed data. In particular, given a candidate model, Mod_C the model estimation component uses parameter estimation techniques to estimate both the time at which the failure occurred, t_F, and the value for the parameters, θ_F, associated with the conjectured failure mode. In this paper we discuss two alternate approaches to our time and parameter estimation problem. The first approach is based on Expectation Maximization (EM) (e.g., (Dempster, Laird, & Rubin 1977)), an iterative technique that converges to an optimal value for t_F and θ_F simultaneously. The second approach we consider employs Generalized Likelihood Ratio (GLR) techniques (e.g., (Basseville & Nikiforov 1993)) to estimate the time of failure t_F, and then uses the observations obtained after the failure to estimate the fault parameters, θ_F, by a least squares regression method. As described in Section 2, the outcome of both approaches is a unique value for t_F and θ_F and a measure of the likelihood of Mod_C given the observations. The proposed approaches to model fitting have trade-offs and we are currently assessing the efficacy of these and other alternative approaches through experimentation.

EM-Based Approach The Expectation Maximization (EM) algorithm (e.g., (Dempster, Laird, & Rubin 1977), (Blimes 1998)) provides a technique for finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given set of data, when that data is incomplete or has missing values. The parameter estimation problem we address in this paper is a variant of the motion segmentation problem described in (Weiss 1997). Here, we define the basic algorithm and the intuition behind our approach. For a more rigorous mathematical account of EM, the reader is referred to (Dempster, Laird, & Rubin 1977).

The time of failure, $t_F = [t_l, t_u]$ of our candidate qualitative diagnosis, $\langle C, \mu_F, t_F, \theta_F \rangle$, dictates the mode in which the failure is conjectured to have oc-

curred. Let us call this mode μ_i. The behavior of our hybrid system in mode μ_i is described by the continuous function f_{μ_i}, with *known* parameters θ_i. At some (to be estimated) time point t_F within the predicted time period of μ_i, we have conjectured that the system experienced a fault which transitions it into mode μ_F. The behavior of our hybrid system in mode μ_F is described by the continuous function f_{μ_F}, with *unknown* parameters, θ_F. We also have a set of data points $\mathcal{O}^{'} = [x_{obs}(0), \ldots, x_{obs}(t)] \subseteq \mathcal{O}$, the observation history, which either reflect the behavior of the system under $f_{\mu i}$ or under f_{μ_F}.

Given all this information, our task is to find 1) values for parameters θ_F, and 2) an assignment of the data points $\mathcal{O}^{'} = [x_{obs}(0), \ldots, x_{obs}(t)]$ to either f_{μ_i} or f_{μ_F} so that we maximize the fit of the data to the two functions. The assignment of data points will in turn tell us the value of t_F. Clearly each assignment is easy given the other. EM provides an iterative algorithm which converges to provide a maximum-likelihood estimate for θ_F given $\mathcal{O}^{'}$, i.e., roughly we are calculating the likelihood of θ, $L(\theta) = P(\mathcal{O}^{'} \mid \theta_F, Mod)$, where Mod, the model, is the sequence of functions $[f_{\mu_i}, f_{\mu_F}]$.

The basic EM algorithm comprises two steps: an Expectation Step (E Step), and a Maximization Step (M Step). The following is a sketch of the algorithm for our task (Weiss 1997):

- Select an initial (random) value for θ_F.
- Iterate until convergence:
  - E Step: assign data points to either $f_{\mu_i}(\theta_i)$ or $f_{\mu_F}(\theta_F)$, which ever fits it best.
  - M Step: re-estimate θ_F using the data points assigned to $f_{\mu_F}(\theta_F)$. θ_F may be estimated using e.g., non-linear regression, depending upon the form of $f_{\mu_F}(\theta_F)$.

We are currently considering several implementations for this algorithm that will exploit problem-specific qualities to help improve convergence of this algorithm. In particular, we may exploit the fact that data points at the end of the $\mathcal{O}^{'}$ sequence must belong to $f_{\mu_F}(\theta_F)$, rather than $f_{\mu_i}(\theta_i)$. Hence we may use these data points to get a better initial estimate of θ_F. Also, we may exploit spatial continuity in the E Step to assign data points to functions. In the general case, EM would assign data points randomly to the two functions. In our case, we know that there is a high likelihood that neighboring data points belong to the same function, and we may exploit this to our advantage.

EM provides a rich algorithm for maximum-likelihood parameter estimation, when we don't know the value of t_F. In some hybrid diagnosis applications, depending upon the sensors in our system, and the level of noise in the sensors, we may be able to develop monitoring techniques that will help isolate a reasonable value for t_F, minimizing the need for iteration in EM. We are beginning to experiment with these techniques to better understand the convergence properties of this technique. We would also like to better understand the mathematical relationship of this technique to alternate approaches.

GLR + Least Squares Approach An alternative to the EM-based approach divides the parameter estimation problem into two parts: (i) estimate the time of failure, t_F, using the Generalized Likelihood Ratio (GLR) method, and (ii) apply a standard least squares method for parameter estimation. The intuition is that solving the problem in two parts simplifies the estimation process, and very likely mitigates the numerical convergence problems that arise in dealing with complex higher-order models.

The GLR method for detecting abrupt changes in continuous signals is described in (Basseville & Nikiforov 1993). We have applied it to fault transients analysis in complex fluid thermal systems (Manders, Mosterman, & Biswas 1999). Here we provide an overview of the method for the single parameter case. Assume that the signal under scrutiny is a time-indexed sequence of random variables $y(k)$, with probability density function, $p_{\theta_i}(y)$ in desired mode μ_i, and $p_{\theta_F}(y)$ in fault mode μ_F. y is either contained in x_{obs} or computed from x_{obs}. We assume that a fault causes an abrupt change in $y(k)$. In the case of the AERCam, y captures the difference between the observed and expected values of the, e.g., acceleration, as predicted by the model.

The central quantity in the change detection algorithm is the cumulative sum of the log-likelihood ratio for a window of observations between times m and n,

$$S_m^n(\theta_F) = \sum_{k=m}^{n} ln \frac{p_{\theta_F}(y(k))}{p_{\theta_i}(y(k))}.$$

Again, this ratio is a function of two unknowns: t_F and θ_F. The common statistical solution is to use maximum likelihood estimates for these two parameters, resulting in a double maximization:

$$g_n = \max_{1 \le m \le n} \sup_{\theta_F} S_m^n(\theta_F).$$

If we assume that probability density functions, $p_{\theta_i}(y)$ and $p_{\theta_F}(y)$ are Gaussian, then g_n reduces to:

$$g_n = \frac{1}{2\sigma_i^2} \max_{1 \le m \le n} \frac{1}{n-m+1} \left[\sum_{k=m}^{n} (y(k) - \omega_i) \right]^2,$$

where ω_i and σ_i^2 are, respectively, the mean and variance for $p_{\theta_i}(y)$.

When processing a sequence of samples, the point of abrupt change, t_F, is computed from $min\{n : g_n \ge h\}$, where h is an appropriately defined threshold. Hence, the smaller the value of h, the more sensitive the function to change, and unfortunately to false alarms, so h must be set carefully.

Once t_F is estimated, data points observed after t_F, are used to estimate the parameter, θ_F for a hypothesized fault using regression techniques. In the case of

the AERCam, the position vector of the AERCam is modeled as a set of quadratic functions in terms of the thruster force. These functions contain one unknown, θ_F, the parameter that corresponds to the degree of degradation in the faulty thruster. The least squares estimate for θ_F is computed, and the the measure of fit of the candidate model to the observed data used to estimated the probability of the candidate diagnosis.

Model Comparison
From the model estimation component, each tracker (see Fig. 4) computes the likelihood of its model Mod_C, and hence of the associated candidate diagnosis $\langle C, \mu_F, t_F, \theta_F \rangle$, as a measure of fit of the observations to the model. As new data $x_{obs}(t)$ are observed, θ_F and t_F, are adjusted and $P(Mod_C \mid x_{obs}(t))$ computed as outlined in Section 2. Different models are compared according to Bayesian model comparison, as described in (MacKay 1991). If the likelihood of Mod_C falls below a predefined acceptable likelihood threshold, α, then its tracker is terminated, and the associated candidate diagnosis $\langle C, \mu_F, t_F, \theta_F \rangle$ removed from the list of candidate diagnoses. Tracking terminates when a unique diagnosis is obtained, or when the diagnoses are sufficiently discriminated to determine suitable controller actions. It is possible that the subsequent actions performed by the controller will not provide the necessary observations to sufficiently discriminate candidate diagnoses. In such cases, active testing must be performed, to discriminate diagnoses. We do not address the issue of active testing in this paper. [3]

5 Discussion and Summary

In this paper we addressed the problem of diagnosing a restricted class of hybrid systems. The main contributions of the paper are 1) formulation of the hybrid diagnosis problem; 2) the exploitation of techniques for qualitative diagnosis of continuous systems to qualitatively reduce the diagnosis search space; and 3) the use of parameter estimation and data fitting techniques for evaluation and comparison of candidate diagnoses.

For computational efficiency, we proposed a simple monitoring and fault detection methodology that was based on flagging individual signal deviations, exceeding a prespecified threshold value. Our implementation will employ a more sophisticated non-linear filtering techniques that ensures a certain number of zero-crossings before a fault is detected. An interesting question that we will have to answer by empirical analysis is whether the GLR method could be employed during the monitoring phase for initial fault detection. The advantage of doing this would be more precise and robust fault detection and time point of failure estimation, that in turn is likely to simplify the candidate generation and parameter estimation process in the model estimation component. The disadvantage is that the GLR technique is computationally expensive, and it is not clear that real-time implementations can be achieved, in the general case. We plan to conduct a number of experimental studies to analyze this issue and related issues concerning the efficacy of alternative time of failure and parameter estimation algorithms.

Clearly the qualitative and quantitative techniques exploited in this paper present only one approach to addressing the problem of diagnosing hybrid systems. Further, our approach was applicable to a restricted class of hybrid systems under some (reasonable) assumptions regarding the nature of faults. In future work we would like to investigate other probabilistic and logic-based techniques (e.g., (Williams & Nayak 1996; McIlraith 1998)) for addressing the problem of diagnosing hybrid systems. We would also like to extend our investigation to a broader class of hybrid systems that include systems whose models contain autonomous jumps (Branicky 1995). In applications where it is relevant, we would like to investigate the role of active testing to support candidate elimination. Finally, our long-term vision is to integrate hybrid diagnosis into a model-based control paradigm that facilitates the continued operation of devices under off-nominal conditions. To do so, diagnosis must be integrated into the action selection process so that diagnosis is performed purposefully to support control decisions.

[3] Note that the technique described here relies on a single-time fault assumption, as observed in Section 2. If multiple independent faults occur in rapid succession this technique may not detect them.

References

Alenius, L., and Gupta, V. 1998. Modeling an AERCam: A case study in modeling with concurrent constraint languages. In *Proceedings of the CP'97 Workshop on Modeling and Computation in the Concurrent Constraint Languages*.

Arnold, V. I. 1978. *Mathematical Methods of Classical Mechanics*. Springer Verlag.

Basseville, M., and Nikiforov, I. 1993. *Detection of Abrupt Changes: Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall.

Blimes, J. A. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, International Computer Science Institute (ICSI) and Computer Science Division, Dept. of Electrical Engineering and Computer Science, U.C. Berkeley.

Branicky, M. 1995. *Studies in Hybrid Systems: Modeling, Analysis, and Control*. Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society Ser. B* 39:1–38.

Etkin, B., and Reid, L. D. 1995. *Dynamics of Flight:Stability and Control*. John Wiley and Sons.

MacKay, D. J. C. 1991. Bayesian interpolation. *Neural Computation* 4(3):415–447.

Manders, E.; Mosterman, P.; and Biswas, G. 1999. Signal to symbol transformation techniques for robust diagnosis in transcend. In *Tenth International Workshop on Principles of Diagnosis*. Submitted for publication.

McIlraith, S. 1998. Explanatory diagnosis: Conjecturing actions to explain observations. In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, 167–177.

Mosterman, P., and Biswas, G. 1997a. Formal specifications for hybrid dynamical systems. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*, 568–573.

Mosterman, P., and Biswas, G. 1997b. Monitoring, prediction, and fault isolation in dynamic physical systems. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, 100–105.

Mosterman, P., and Biswas, G. 1999a. Building hybrid observers for complex dynamic systems using model abstractions. In *International Workshop on Hybrid Systems: Computation and Control*.

Mosterman, P., and Biswas, G. 1999b. Diagnosis of continuous valued systems in transient operating regions. *IEEE Transactions on Systems, Man, and Cybernetics*. To appear.

Rinner, B., and Kuipers, B. 1999. Monitoring piecewise continuous behaviour by refining trackers and models. In *AAAI Technical Report – AAAI 1999 Spring Symposium on Hybrid Systems and AI*. To appear.

Sweet, W. 1995. The glass cockpit. *IEEE Spectrum* 30–38.

Weiss, Y. 1997. Motion segmentation using EM – a short tutorial. http://www-bcs.mit.edu/people/yweiss/tutorials.html.

Williams, B., and Nayak, P. 1996. A model-based approach to reactive self-configuring systems. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, 971–978.