Model-based health monitoring for a vehicle steering system with multiple faults of unknown types

Yu, Ming; Wang, Danwei

2013

Yu, M., & Wang, D. (2014). Model-Based Health Monitoring for a Vehicle Steering System With Multiple Faults of Unknown Types. IEEE Transactions on Industrial Electronics, 61(7), 3574-3586.

https://hdl.handle.net/10356/103915

https://doi.org/10.1109/TIE.2013.2281159

Model-Based Health Monitoring for a Vehicle Steering System With Multiple Faults of Unknown Types

Ming Yu, *Member, IEEE*, and Danwei Wang, *Senior Member, IEEE*

***Abstract*—This paper presents a model-based fault diagnosis and prognosis scheme for a vehicle steering system. The steering system is modeled as a hybrid system with continuous dynamics and discrete modes using the hybrid bond graph tool. Multiple faults of different types, i.e., abrupt fault, incipient fault, and intermittent fault, are considered using the concept of Augmented Global Analytical Redundancy Relations (AGARRs). A fault discriminator is constructed to distinguish the type of faults once they are detected. After that, a fault identification scheme is proposed to estimate the magnitude of abrupt faults, the characteristic of intermittent faults, and the degradation behavior of incipient faults. The fault identification is realized by using a new adaptive hybrid differential evolution (AHDE) algorithm with less control parameters. Based on the identified degradation behavior of incipient faults, prognosis is carried out to predict the remaining useful life of faulty components. The proposed algorithm is verified experimentally on the steering system of a CyCab electric vehicle.**

***Index Terms*—Adaptive hybrid differential evolution (AHDE), augmented global analytical redundancy relations (AGARRs), fault discriminator, model-based diagnosis and prognosis, remaining useful life (RUL).**

Nomenclature

AGA	Adaptive genetic algorithm
AGARRs	Augmented global analytical redundancy relations
AHDE	Adaptive hybrid differential evolution
ARRs	Analytical redundancy relations
BDE	Binary-valued differential evolution
BG	Bond graph
DC	Direct current
DE	Differential evolution
DHBG	Diagnostic hybrid bond graph
EOL	End of life
FDI	Fault detection and isolation
FDV	Fault discrimination vector
FSM	Fault signature matrix
GA	Genetic algorithm
GARRs	Global analytical redundancy relations
HBG	Hybrid bond graph
HDE	Hybrid differential evolution
MCSM	Mode change signature matrix
MD-FSM	Mode-dependent fault signature matrix
ODE	Ordinary differential equation
PCI	Peripheral component interconnect
PWM	Pulse-width modulation
RDE	Real-valued differential evolution
RUL	Remaining useful life
SCAPH	Sequential causality assignment procedure for hybrid systems

I. Introduction

FAULT diagnosis of hybrid systems is an active area of research in recent years [2]–[4]. Generally speaking, a hybrid system consists of interacting continuous dynamics and discrete behaviors. Discrete behaviors are represented by modes. Hybrid systems emerge from manufacturing system, automotive engine control, chemical process, aerospace engineering as well as embedded control system. Health monitoring of hybrid systems requires information from continuous part as well as discrete part. In hybrid system diagnosis, both parametric fault and fault mode can affect current system behavior. A parametric fault refers to the deviation of parameter value from its nominal value to an unknown one. For example, a flat tire fault in a vehicle will increase the friction coefficient between ground and tire. For a fault mode, the faulty state is known *a priori* and can be modeled by known parameters only. Two examples are short-circuits and open-circuits faults of power switches. In addition, unpredictable mode changes may occur at any time. All these factors add to the complexity of fault diagnosis in hybrid systems.

In general, model-based diagnosis methods provide satisfactory performance while requiring the development of accurate mathematical model to describe the system under monitoring. BG has been successfully exploited for model-based FDI due to its capability of modeling complex systems in a unified way [5]. Moreover, the structural properties deduced from the causalities on the graph can be utilized for systematic generation of ARRs for both linear and nonlinear systems. The resultant ARRs only contain known parameters and measurements which can be used for online fault detection as well as offline monitoring ability analysis. Since the physical components are clearly presented on the graph model, BG is also applied to optimize sensor placement [5].

The development of HBG facilitates the modeling of hybrid systems using the BG technique with the help of controlled

Manuscript received October 29, 2012; revised March 26, 2013 and June 10, 2013; accepted August 13, 2013. Date of publication September 9, 2013; date of current version January 31, 2014.

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yuming@ntu.edu.sg; edwwang@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2013.2281159

junctions [16]. Based on the HBG, several issues about FDI and mode tracking technique for hybrid systems are discussed in [3], [4]. BG-based FDI schemes for vehicle systems are reported in [2], [10], and [15]. In [2], an electro-hydraulic steering system of an electric vehicle and faults is modeled as a hybrid dynamic system by the HBG modeling technique. This method is based on the global constraint called GARRs to carry out fault detection, isolation, and estimation. However, this method addresses only single fault scenario and does not consider prognosis for incipient faults. In [10], robust FDI of an electric vehicle with structured and unstructured uncertainties is developed. The generation of a nonlinear model and residuals for the vehicle system with adaptive thresholds is synthesized using BG tool and linear fractional transformation. A super-twisting observer is used to estimate both unstructured uncertainties and unknown inputs. Similarly, residual generation for actuators FDI of an electric vehicle is proposed in [15]. The modeling step is accomplished using the BG theory. Unmeasured flow variables and system nonlinearities are estimated by a nonlinear observer with finite time convergence and are considered as unknown inputs in BG model. The FDI methods developed in [10] and [15] are based on continuous dynamics and have no design issues about prognosis.

Failure prognosis, on the other hand, tries to determine whether a failure is impending and estimates how soon and how likely a failure will occur. The main task of prognosis is to predict the EOL or RUL of a faulty component or subsystem [9]. Diagnosis and prognosis are two important aspects in condition-based maintenance, and they are complementary tasks since diagnosis is a "static" indicator whereas prognosis is a "dynamic" indicator. In [13], a blind deconvolution denoising scheme is developed and applied to vibration signals collected from a test-bed of the helicopter main gearbox subjected to a seeded fault. The quality of the features is improved using the proposed denoising scheme, and hence the failure prognostic algorithm is more accurate. However, in some industrial applications, vibration signals are not readily available. An interacting multiple model-based prognosis method is proposed in [11] to track the hidden damage. Remaining-life prediction is performed by mixing mode-based life predictions via time-averaged mode probabilities.

In this paper, a quantitative HBG model-based fault diagnosis and prognosis method is developed for the steering system of an electric vehicle. The system is mixed with various types of fault, including abrupt fault, intermittent fault, incipient fault, and fault mode. The fault type for any fault is unknown beforehand and a fault discriminator is constructed to distinguish the types of faults once they are detected. A new AHDE algorithm is proposed to realize the fault identification task and then provide the information about the magnitude of abrupt faults, the characteristic of intermittent faults, and the degradation behavior of incipient faults. A failure prognosis scheme is developed for the incipient faults with identified degradation behaviors.

The main contributions of this paper are threefold.

1) A single framework, which considers FDI of multiple abrupt, incipient, intermittent faults, and fault modes (i.e., faults represented by modes) as well as prognosis of incipient faults, is developed for health monitoring of hybrid systems.
2) A method for fault modeling and coding, which allows developments of a fault discriminator and enables estimations of fault characteristics (e.g., for intermittent fault, the magnitude and frequency of repetitions), is proposed.
3) A new DE search algorithm called AHDE is developed. The hybrid comes from its ability of finding real-valued solutions (with RDE) together with binary-valued solutions (with BDE). An adaptive scheme for selecting some of the algorithm control parameters based on an improvement measure is proposed (meaning these control parameters are automatically tuned by the algorithm). As a result, the total number of control parameters (that are required to be tuned manually) is decreased.

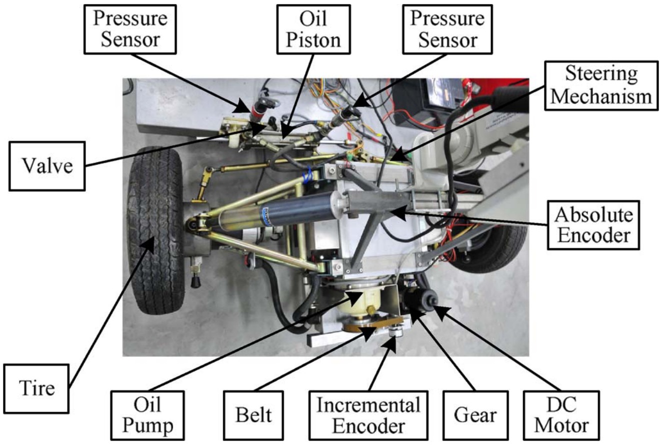

Fig. 1. Vehicle steering system.

This paper is organized as follows: In Section II, the model of the vehicle steering system is presented. Section III describes the proposed diagnosis and prognosis approaches, then, in Section IV, experimental results are reported. Finally, Section V concludes the paper.

II. Vehicle Steering System and Its Model

A. Electro-Hydraulic Steering System

This section studies the electro-hydraulic steering system from the considered electric vehicle. The steering system, as shown in Fig. 1, is composed of dc motor, gear, belt, oil pump, oil piston, steering mechanism, and tires. There are totally four sensors, i.e., two pressure sensors, an absolute encoder, and an incremental encoder, mounted on the system. Two pressure sensors are installed on both sides of the hydraulic cylinder to measure the pressure of the two chambers. The absolute encoder is located under the platform to record the steering angle. The output speed of the dc motor is measured by the incremental encoder mounted on the gear output shaft. The steering power is provided from the hydraulic actuator which generates the pressure difference to push the oil piston inside the actuator. The oil piston is connected to the steering mechanism. The oil is pumped from an internal oil reservoir to the hydraulic actuator through an oil tube and can be returned to the pump through another oil tube. The pump is driven by dc motor through the belt and converts mechanical speed to proportional oil flow.

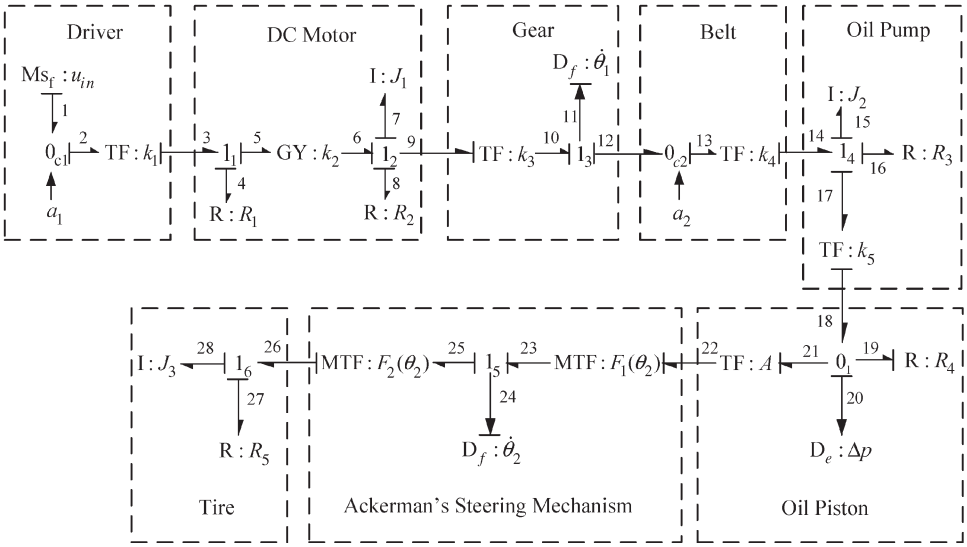

Fig. 2. DHBG model of the steering system.

B. Modeling of the System Using the HBG Technique

The DHBG of the electro-hydraulic steering system is given in Fig. 2. The DHBG, developed in [4], is an HBG model equipped with a specific causality assignment procedure, i.e., SCAPH introduced in [3]. The objective of the DHBG is to avoid the causality reassignment after mode change happens in HBG model [16]. For a given acausal HBG, i.e., HBG without causality assignment, it is desirable to assign a causality to the HBG such that all controlled junctions are in preferred causalities. The purpose of the preferred causality is to restrict the output variable of the controlled junction as the input variable of those components which pose no invalid causal form to the HBG during the off state of the controlled junction. Thus, a unified description of HBG can be obtained for all modes using the concept of DHBG.

The motor driver is composed of the input source $MSf : u_{in}$ and the voltage-to-current constant k_1 which is modeled by BG element $TF : k_1$. The controlled junction 0_{c1} is adopted to model a burnt driver fault or a burnt motor fault. The state variable a_1 of junction 0_{c1} is 0, i.e., $a_1 = 0$, when a fault occurs. The dc motor consists of electrical part and mechanical part. The electrical part is modeled by electrical resistance $R : R_1$ and gyrator $GY : k_2$. The BG element GY is used to describe that the torque generated is linear to the current with current-to-torque ratio k_2. The mechanical part of the dc motor is modeled by the motor inertia $I : J_1$ and mechanical friction $R : R_2$. The mechanical friction includes the viscous friction coefficient k_{f_2} and Coulomb friction coefficient Fu_2.

The transmission from the motor to the pump is modeled by the gear ratio $TF : k_3$ and the belt ratio $TF : k_4$. An incremental sensor, modeled by BG sensor element $Df : \dot{\theta}_1$, is used to measure the velocity $\dot{\theta}_1$ at the gear output shaft. A state variable a_2 of junction 0_{c2} is used to model the broken belt fault and $a_2 = 0$ when the belt is broken. Thus, the pump rotor velocity is zero after the fault happens. The rotor of the pump is composed of mechanical friction $R : R_3$ with coefficients k_{f_3} and Fu_3, and rotor inertia $I : J_2$. The transformer element $TF : k_5$ describes the linear relation between pump angular velocity and oil flow through the pump. The oil piston inside the hydraulic actuator is modeled by a transformer $TF : A$ with A representing the effective cross area of the piston. To model the internal leakage of the piston, resistance element $R : R_4$ is utilized where R_4 denotes the internal resistance to oil flow. The following nonlinear relation is used to describe the dynamics of the leakage [5]:

$$f_{19} = \frac{\mathrm{sign}(e_{19})\sqrt{|e_{19}|}}{R_4} = \frac{\mathrm{sign}(\Delta p)\sqrt{|\Delta p|}}{R_4} \tag{1}$$

where $\Delta p = p_1 - p_2$ is the pressure difference of the two pressure sensors and $\mathrm{sign}(\cdot)$ represents the sign function.

When the piston is normal, then $R_4 \to \infty$, and $R_4 \ll \infty$ under leakage condition. The steering mechanism is modeled by $F_1(\theta_2)$ and $F_2(\theta_2)$ which are derived from the Ackerman's geometry. The absolute encoder is modeled by $Df : \dot{\theta}_2$ which measures the steering angle θ_2. The tire is modeled by the inertial $I : J_3$ and the friction between road and tire $R : R_5$, including coefficients k_{f_5} and Fu_5.

III. Diagnosis and Prognosis Approaches

A. Residual Generation

To carry out FDI analysis, all controlled junctions and storage elements of the DBHG are assigned preferred causality as shown in Fig. 2. The sensors on the DHBG are assigned inverted causality, and the constitutive relation of each sensor junction is utilized to develop an AGARR. GARRs describe the behaviors of a hybrid system at all operating modes [4]. However, these constraints cannot be utilized for the identification of fault in nonparametric components, such as sensors and actuators. The development of AGARR extends the capability of GARRs to identify the degradation of components, such as sensors and actuators, which cannot be described by physical parameters [12]. In each AGARR, sensor measurement or source element is associated with an efficiency factor β to quantify the severity of the fault in the nonparametric component. When the system is normal, all β are equal to one and in case of faults, those β related to fault candidates become unknown parameters.

It is worth to note that AGARR extends GARR by adding an efficiency factor to any input or measurement. This idea is effective for the case of multiplicative faults because the augmentation is based on a multiplicative structure. In some cases, such as additive faults (e.g., a bias, which is a very common fault for sensors), it may be more efficient if a constant additive fault can be identified as a constant additive for fault size identification. Nevertheless, the multiplicative structure through the introduction of the concept of the efficiency factor is more efficient for prognosis purpose in which typically a time varying fault representation is required. For example, to model partial loss of the control effectiveness in the actuator, a time varying efficiency factor with multiplicative representation, instead of a constant additive fault, could be more efficient for prognosis of the actuator.

First, consider the constitutive relation of junction 1_3 attached with flow sensor $Df : \dot{\theta}_1$

$$e_{11} = e_{10} - e_{12} = 0. \tag{2}$$

Based on the causal paths on DHBG, the unknown variables e_{10} and e_{12} can be expressed as

$$\begin{aligned} e_{10} &= k_3^{-1} e_9 = k_3^{-1}(e_6 - e_7 - e_8) \\ &= k_3^{-1}\left(k_2 f_5 - J_1 \dot{f}_7 - k_{f_2} f_8 - Fu_2 \mathrm{sign}(f_8)\right) \\ &= k_3^{-1}\Big(a_1 k_1 k_2 u_{in} - J_1 k_3^{-1} \ddot{\theta}_1 \\ &\qquad -k_{f_2} k_3^{-1} \dot{\theta}_1 - Fu_2 \mathrm{sign}(\dot{\theta}_1)\Big) \end{aligned} \tag{3}$$

$$\begin{aligned} e_{12} &= a_2 e_{13} = a_2 k_4 e_{14} = a_2 k_4 (e_{15} + e_{16} + e_{17}) \\ &= a_2 k_4 \left(J_2 \dot{f}_{15} + k_{f_3} f_{16} + Fu_3 \mathrm{sign}(f_{16}) + k_5 e_{18}\right) \\ &= a_2 k_4 \Big(J_2 k_4 \ddot{\theta}_1 + k_{f_3} k_4 \dot{\theta}_1 \\ &\qquad +Fu_3 \mathrm{sign}(\dot{\theta}_1) + k_5 \Delta p\Big). \end{aligned} \tag{4}$$

Substituting (3) and (4) into (2) gives

$$\begin{aligned} &\mathrm{AGARR}_1 \\ &= k_3^{-1}\Bigg(a_1 k_1 k_2 u_{in} - J_1 k_3^{-1} \frac{d^2}{dt^2}\left(\frac{\theta_1}{\beta_{\theta_1}}\right) \\ &\qquad -k_{f_2} k_3^{-1} \frac{d}{dt}\left(\frac{\theta_1}{\beta_{\theta_1}}\right) - Fu_2 \mathrm{sign}\left(\frac{d}{dt}\left(\frac{\theta_1}{\beta_{\theta_1}}\right)\right)\Bigg) \\ &\quad - a_2 k_4 \Bigg(J_2 k_4 \frac{d^2}{dt^2}\left(\frac{\theta_1}{\beta_{\theta_1}}\right) + k_{f_3} k_4 \frac{d}{dt}\left(\frac{\theta_1}{\beta_{\theta_1}}\right) \\ &\qquad +Fu_3 \mathrm{sign}\left(\frac{d}{dt}\left(\frac{\theta_1}{\beta_{\theta_1}}\right)\right) + k_5 \frac{\Delta p}{\beta_{\Delta p}}\Bigg) = 0. \end{aligned} \tag{5}$$

Next, consider junction 0_1 with sensor $De : \Delta p$; the constitutive relation of this junction can be expressed as

$$f_{20} = f_{18} - f_{19} - f_{21} = 0. \tag{6}$$

The flow variables f_{18} and f_{21} can be represented as

$$f_{18} = k_5 f_{17} = a_2 k_4 k_5 \dot{\theta}_1 \tag{7}$$

$$f_{21} = A f_{22} = A F_1^{-1}(\theta_2)\dot{\theta}_2. \tag{8}$$

Combining (6)–(8) and (1), AGARR_2 can be expressed as

$$\begin{aligned} \mathrm{AGARR}_2 = a_2 k_4 k_5 \frac{d}{dt}\left(\frac{\theta_1}{\beta_{\theta_1}}\right) - \frac{\mathrm{sign}\left(\frac{\Delta p}{\beta_{\Delta p}}\right)\sqrt{\left|\frac{\Delta p}{\beta_{\Delta p}}\right|}}{R_4} \\ -AF_1^{-1}\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{d}{dt}\left(\frac{\theta_2}{\beta_{\theta_2}}\right) = 0. \end{aligned} \tag{9}$$

Finally, consider junction 1_5 attached with flow sensor $Df : \dot{\theta}_2$

$$e_{24} = e_{23} - e_{25} = 0. \tag{10}$$

Unknown variables e_{23} and e_{25} can be solved as

$$e_{23} = F_1^{-1}(\theta_2) e_{22} = AF_1^{-1}(\theta_2)\Delta p \tag{11}$$

$$\begin{aligned} e_{25} &= F_2(\theta_2) e_{26} = F_2(\theta_2)(e_{27} + e_{28}) \\ &= F_2(\theta_2)\left(J_3 \dot{f}_{28} + k_{f_5} f_{27} + Fu_5 \mathrm{sign}(f_{27})\right) \\ &= F_2(\theta_2)\Bigg(J_3 \frac{d}{dt}\left(F_2(\theta_2)\dot{\theta}_2\right) + k_{f_5} F_2(\theta_2)\dot{\theta}_2 \\ &\qquad + Fu_5 \mathrm{sign}\left(F_2(\theta_2)\dot{\theta}_2\right)\frac{d}{dt}\Bigg). \end{aligned} \tag{12}$$

The third AGARR can be obtained according to (10)–(12)

$$\begin{aligned} &\mathrm{AGARR}_3 \\ &\quad = AF_1^{-1}\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{\Delta p}{\beta_{\Delta p}} \\ &\qquad - J_3 F_2\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{d}{dt}\left[F_2\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{d}{dt}\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\right] \\ &\qquad - k_{f_5} F_2^2\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{d}{dt}\left(\frac{\theta_2}{\beta_{\theta_2}}\right) \\ &\qquad - Fu_5 F_2\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\mathrm{sign}\left[F_2\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\frac{d}{dt}\left(\frac{\theta_2}{\beta_{\theta_2}}\right)\right] = 0. \end{aligned} \tag{13}$$

B. Considered Types of Fault

In the vehicle steering system, various sources of faults are considered. These sources of faults include parametric faults (internal leakage of hydraulic cylinder and flat tire), sensor faults (pressure sensor fault, incremental encoder fault, and absolute encoder fault), and fault modes (broken belt, burnt driver, or burnt motor). For the parametric faults (usually, components are represented by their physical parameters in the dynamic model), a deviation of parameter value from its nominal one is expected upon the fault occurrence. For example, a flat tire fault will increase the friction coefficient between ground and tire.

For the sensor faults, usually there are discrepancies between measured signals and their actual values. For instance, due to a wire cutoff fault in the incremental encoder, pulses are not generated by the encoder and the counter reading at the encoder interface card remains constant. As a result, the measured velocity is read as zero, although the system is still running. A fault mode is defined if the faulty state can be modeled as a system mode, i.e., all parameters at the faulty state are known a priori. Few examples are burnt motor and broken transmission belt.

The nature of possible faulty situations in the steering system consists of three types, i.e., abrupt faults, incipient faults, and intermittent faults. Abrupt faults are typically modeled as step-like deviation in which the component value is abruptly changed from its nominal value to an unknown faulty one. Incipient faults are slowly developing and intermittent fault usually manifests itself intermittently in an unpredictable manner. Basically, abrupt faults and incipient faults belong to persistent faults, which means that once they appear, do not disappear, while intermittent faults do. For the parametric faults and sensor faults in the steering system, both can be modeled using different natures (abrupt, incipient, and intermittent). For example, the internal leakage of hydraulic cylinder can be modeled as an abrupt fault due to a leakage hole inside the cylinder, and it could be also considered as an incipient fault due to the slowly developed worn seal. However, for the fault mode, the fault nature is abrupt in which the states of the controlled junction (i.e., a_1 for burnt motor and a_2 for broken belt) change instantaneously.

For the abrupt fault, the component value is abruptly changed from its nominal value to an unknown faulty value. This fault profile $\overline{P}_{ab}$ can be represented as

$$\overline{P}_{ab} = \begin{cases} P & \text{if } t < t_0 \\ P_{ab} & \text{if } t \geq t_0 \end{cases} \tag{14}$$

where P represents the nominal value of component or efficient factor, P_{ab} denotes the unknown faulty value, and t_0 represents the fault occurrence time.

Intermittent fault usually manifests itself intermittently in an unpredictable manner. In some situation, it is possible that the same type of fault repeats multiple times at different intervals [14]. Similarly, intermittent or nonpersistent faults may occur repeatedly. For instance, at a bottle filling station, a multiple number of bottles may not be filled properly at different time intervals. For intermittent faults, it is difficult to handle due to hard anticipation. Early detection of these faults might give an indication of when maintenance is required, minimizing the probability of system/component failure [5]. It is beneficial to develop a method which is capable of early detecting these faults, and at the same time evaluating their severity and frequency. In this paper, assume that the intermittent fault is periodic with fixed faulty value. Thus, the fault profile of intermittent fault $\overline{P}_{int}$ can be expressed as

$$\overline{P}_{int} = -\frac{P - P_{int}}{2}\text{sign}\left\{\sin\left[\frac{2\pi(t-t_0)}{T}\right]\right\} + \frac{P + P_{int}}{2} \tag{15}$$

where P_{int} denotes the unknown faulty value, and T denotes the period of the intermittent fault.

Incipient faults are usually related to the wear and tear of the system components. It is relatively difficult to detect the incipient fault due to slowly developing nature of the fault and the compensation effect of feedback control. Incipient failures are more important where it is required that slowly developing fault is detected early enough to avoid more serious consequences. The fault profile $\overline{P}_{inc}$ related to incipient fault is represented as

$$\dot{\overline{P}}_{inc} = b_1\overline{P}_{inc} \tag{16}$$

$$\dot{\overline{P}}_{inc} = b_2\overline{P}_{inc}^2 \tag{17}$$

$$\ddot{\overline{P}}_{inc} = b_3\dot{\overline{P}}_{inc} + \overline{P}_{inc} + b_4 \tag{18}$$

$$\ddot{\overline{P}}_{inc} = b_5\dot{\overline{P}}_{inc}^2 + b_6\overline{P}_{inc} + b_7 \tag{19}$$

where b_1, b_2, b_3, b_4, b_5, b_6, and b_7 are unknown degradation model coefficients, and $\overline{P}_{inc}(t_0) = P$.

Equations (16)–(19) are able to describe different dynamic behaviors, linear or nonlinear, related to the incipient fault with various model coefficients. In this paper, the exact degradation model, i.e., one of the ODE equations in (16)–(19), for any particular fault is unknown in advance. Note that these four incipient fault profiles are only examples of four possible degradation models and others can be considered as well if more information is given on the nature of the fault.

C. Diagnosis of Multiple Faults of Unknown Types

In some actual complex systems, faults can occur both on the mechanical part and on the electrical part of the system. The actuator and sensor faults are more frequent due to the presence of electrical parts and connections [1]. Fault diagnostic algorithm considering only parametric faults is not sufficient for real application. In addition, it is difficult to obtain the information of fault type for a particular fault in most cases. Therefore, it is desirable to develop a method that is able to handle fault diagnosis and prognosis when the knowledge of fault type is unavailable.

It is worth to note that once a fault mode occurs, i.e., burnt driver or broken belt, the vehicle system is completely failed, it is not meaningful, under this condition, to consider other fault types in parametric or nonparametric components. In addition, if the burnt driver (fault mode) occurs, the whole steering system is not moving even though the command input is not zero, which means that there is no signal flow in the system. As a result, no other fault could be detected under this condition. Therefore, in this paper, it is assumed that when a fault mode occurs, other components are normal. Using this assumption, the mode $[a_2\ a_1] = [1\ 1]$ is the only mode where the multiple component faults may occur. There are seven possible faults, i.e., two parametric faults, three sensor faults, and two fault modes, considered in this work.

Table I is a MD-FSM which represents the cause–effect relations between component faults (parametric and nonparametric) and residuals (numerical evaluation of AGARRs) under different operating modes. The column headers in MD-FSM represent the residuals and fault detectability (D_b). Each entry

TABLE I
MD-FSM at Mode $[a_1\ a_2] = [1\ 1]$

	AGARR$_1$	AGARR$_2$	AGARR$_3$	D_b
R_4	0	1	0	1
R_5	0	0	1	1
β_{θ_1}	1	1	0	1
β_{θ_2}	0	1	1	1
$\beta_{\Delta p}$	1	1	1	1

TABLE II
MCSM of the Steering System

	AGARR$_1$	AGARR$_2$	AGARR$_3$	D_b	I_b
a_1	1	0	0	1	1
a_2	1	1	0	1	1

of the table holds a Boolean value. A "1" in an entry indicates that the residual under the column is sensitive to the fault of the corresponding component that lies in the matching row which means that the component appears in the AGARR equation. On the contrary, a "0" in the entry represents that the residual is insensitive to the fault of the component which indicates that the AGARR equation does not include this component. For each row, the entries under the columns form the fault signature of the component. If at least a 1 appears in the fault signature of the parameter, then the component fault is said to be fault detectable. This ability is represented by $D_b = 1$ in the matrix. Since for a hybrid system, different mode changes may have different influences on the system's residuals; these influences are represented by a matrix, called MCSM. Table II is an MCSM which represents cause–effect relations between mode changes and residuals. A "1" in an entry indicates that the residual under the column is sensitive to the mode change of the corresponding controlled junction that lies in the matching row. A "0"- in the entry represents that the residual is insensitive to the mode change of the corresponding controlled junction. Mode change detection ability and mode change isolation ability are presented in the last two columns of the MCSM. If the mode change is detectable by at least one AGARR, then $D_b = 1$; if the mode change signature is unique, then the mode change is isolable and $I_b = 1$. Both MD-FSM and MCSM are developed offline based on the AGARR equations. Note that in Table I, fault isolability denoted by I_b is not involved because even if the fault signature of a component is unique in the table, the coherence vector observed which matches this fault signature may be caused by several faults occurring simultaneously.

The fault detection is carried out by online evaluating the residuals, and a fault is detected if one of the residuals crosses the predefined threshold. A coherence vector $C = [c_1\ c_2\ c_3]$ is established to show residuals inconsistency in which $c_i = 1$, $i = 1, 2, 3$, if AGARR$_i$ is inconsistent (i.e., exceeds the corresponding threshold) and zero otherwise. When the system is fault free, the binary coherence vector C will be zero. On the other hand, if the system is faulty, then at least one entry of the coherence vector will take value 1. Based on the detected coherence vector, a set of fault candidates is established. If the set of fault candidates includes a suspected fault mode, then a faulty mode estimator is first activated and if no new fault mode is identified, i.e., the identified mode is $[a_2\ a_1] = [1\ 1]$, then component fault estimator is activated. Since the fault type of any fault is unknown before hand, a fault discriminator is constructed to determine the type, i.e., abrupt, incipient, or intermittent, of the component fault. For the fault discriminator, a binary vector $\zeta = [\zeta_{ab}\ \zeta_{int}\ \zeta_{inc}]$, called FDV, is defined and is related to one fault candidate in the fault set. The fault type for any fault can be determined by FDV during the identification process. Thus, $\zeta = [1\ 0\ 0]$ represents abrupt fault, $\zeta = [0\ 1\ 0]$ denotes intermittent fault, and $\zeta = [0\ 0\ 1]$ represents incipient fault. It is obvious that during the identification process, condition $\zeta_{ab} + \zeta_{int} + \zeta_{inc} = 1$ should be satisfied. As for the incipient fault, the exact degradation behavior, i.e., one of the equations from (16)–(19), is unknown beforehand; thus, it is required to find the true degradation model together with model coefficients. To solve this problem, if the identified FDV $\zeta = [0\ 0\ 1]$, then for each incipient fault, a model selection scheme is developed as follows. A binary vector is defined as $\eta = [\eta_1\ \eta_2]$ and let the four binary values associate with the four prescribed models, where $\eta = [0\ 0]$ means that the degradation model fulfills the equation in (16); $\eta = [0\ 1]$, $\eta = [1\ 0]$, and $\eta = [1\ 1]$ indicate that the degradation model matches (17)–(19), respectively. During the identification process, if the identified fault discriminator vector indicates that this fault is incipient nature, then the binary vector η and the degradation model coefficients will be estimated together; those results which can maximize the fitness function in (24) can be considered as true degradation model with model coefficients. In this way, the model selection is represented by the binary vector selection in the identification process. The proposed identification method is illustrated in Fig. 3.

The formula in (15) is essentially a discontinuous function, which might lead to problem if conventionally analytical methods, such as gradient-based methods, are used for the intermittent fault period identification considering the non-differentiable function. In addition, the FDV related to each fault is not involved in the fitness function; thus, no gradient information can be obtained related to the binary numbers. Therefore, heuristics-based DE can be an alternative. DE is invented by Storn and Price and is able to deal with non-differentiable, nonlinear, and multimodal objective functions [6]. Many published works demonstrated that DE converges fast and is robust, simple in implementation, and requires only a few control parameters [8]. From the aforementioned analysis, it is known that the solution for the identification problem includes real numbers for fault values for abrupt fault and intermittent fault, intermittent fault period, and degradation model coefficients for incipient fault, as well as binary numbers for FDV and η. An HDE algorithm is developed, in which RDE is adopted to find the real numbers and BDE is utilized to search the binary numbers.

Assume the solution space is D-dimensional and the ith individual in RDE is represented by a vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$. During mutation at generation G, two individuals $x_{r_1}^G$ and $x_{r_2}^G$ satisfying $r_1 \neq r_2$ and $r_1, r_2 \in [1, M]$ in population are randomly selected. Using strategy DE/best/1/bin, the mutation operator can be represented as

$$V_{ij}^{G+1} = X_{best}^G + F^r \times \left(X_{r_1 j}^G - X_{r_2 j}^G\right) \tag{20}$$

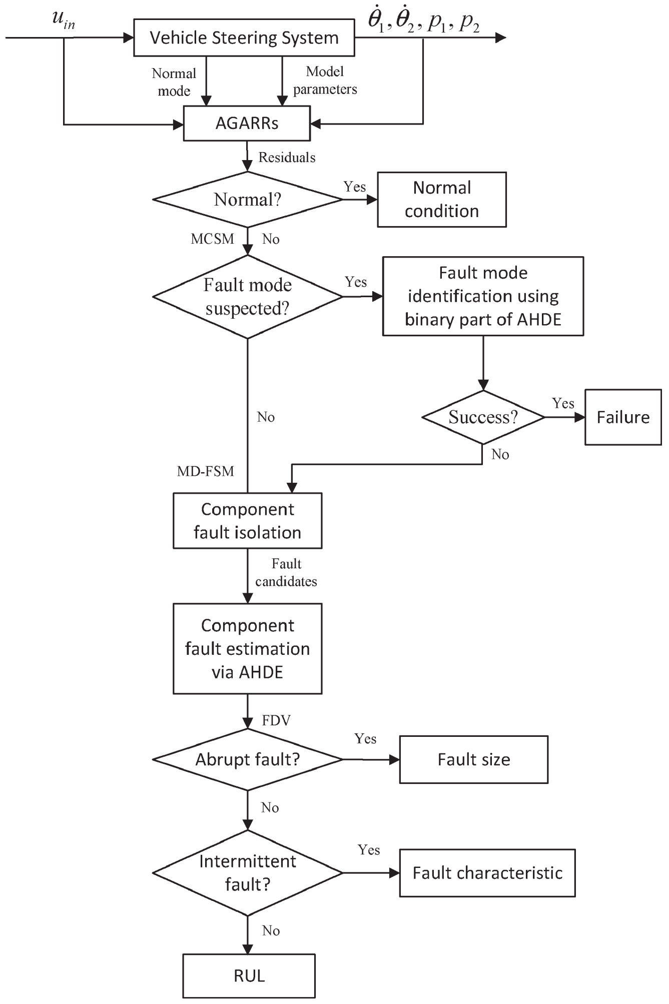

Fig. 3. Proposed health monitoring scheme.

where F^r is the differential factor satisfying $F^r \in [0, 2]$ and X_{best}^G is the best individual in population at generation G.

In crossover

$$H_{ij}^{G+1} = \begin{cases} V_{ij}^{G+1} & \text{if } rand_j \leq CR^r \text{ or } j = rnbr(i) \\ X_{ij}^G & \text{if } rand_j > CR^r \text{ or } j \neq rnbr(i) \end{cases} \quad (21)$$

where $CR^r \in [0, 1]$ is the crossover factor, and $rnbr(i)$ is a randomly selected index from D dimensions to ensure that at least one dimension parameter from V_i^{G+1} can be obtained by H_i^{G+1}[6].

Selection operates by comparing the individuals' fitness to generate the new population of next generation

$$X_i^{G+1} = \begin{cases} H_i^{G+1} & \text{if } f\left(H_i^{G+1}\right) > f\left(X_i^G\right) \\ X_i^G & \text{otherwise.} \end{cases} \quad (22)$$

In BDE, the solution is represented by a binary string instead of a real value vector [7]. Since crossover operates as discrete form in continuous space, it can be directly used in BDE. Mutation is the only operator which should be modified

$$V_{ij}^{G+1} = X_{best}^G + F^b \bullet \left(X_{r_1 j}^G \oplus X_{r_2 j}^G\right) \quad (23)$$

where symbols $(\bullet)$, $(+)$, and $(\oplus)$ denote Boolean algebra "AND," "OR," and "XOR," respectively. F^b is a random D-bit binary vector and it is not a control parameter.

The RDE and BDE evolve simultaneously and are coupled through the common fitness function

$$F_{fitness} = 1/\left(\sum_{l=1}^{m}\sum_{n=1}^{N}|G_l^n| + \epsilon\right) \quad (24)$$

where G_l is the lth AGARR equation, ϵ is a small positive constant which is used to avoid zero division during the search process, n is the discrete sampling index.

In this paper, an AHDE algorithm is developed, in which an adaptive scheme for selecting control parameters CR^r and F^r in RDE, and CR^b (crossover factor for BDE) in BDE is proposed as follows:

$$CR^r = \sin\left(\frac{\pi}{2}IM\right) \cdot rand \quad (25)$$

$$CR^b = \sin\left(\frac{\pi}{2}IM\right) \cdot rand \quad (26)$$

$$F^r = 2\sin\left(\frac{\pi}{2}IM\right) \cdot rand \quad (27)$$

where IM is the improvement measure and is defined as

$$IM = \frac{F_{fitness}^{best,G-1}}{F_{fitness}^{best,G}}. \quad (28)$$

The purpose of the adaptive scheme described in (25)–(27) is to dynamically tune the key parameters of HDE based on the improvement measure in (28). When IM is close to one, it indicates that the solution from the Gth generation has an insufficient improvement over the one from the previous generation $G - 1$. It is desirable to increase its search space, and thus set F^r with range [0 2] and choose CR^r and CR^b with range [0 1]. On the other hand, when IM is close to zero, it means that the algorithm obtains a a significant improvement between two successive generations. To provide a finer search, a smaller range $[0\ \sin((\pi/2)IM)]$ is chosen for CR^r and CR^b, and $[0\ 2\sin((\pi/2)IM)]$ for F^r. As a result, the total number of the control parameters for the HDE algorithm is decreased.

If the number of the fault candidates in the fault set is more than one, several AHDE estimators are required to run in parallel. The number of the AHDE estimators is equal to the number of fault candidates in the fault set. Finally, the results from the AHDE estimator with the highest fitness value are considered to be best estimated results. As for the fault mode identification, since the state variables of the mode are binary numbers, only part of AHDE is used. This part of AHDE contains BDE with adaptive law described in (26). The identified mode is compared with the previous known mode; if a new fault mode is identified, the system is completely failed.

It is worthy to note that for the incipient fault, the fault parameter value is not changed at the instance of fault occurrence but it slowly increases or decreases over time. Thus, the residual containing such incipient fault cannot detect the fault at the time of fault occurrence. Moreover, due to the slowly developing nature of the incipient fault, the small effects of this fault on residuals in the beginning stage might be masked by the effects of measurement noise. As a result, a delay exists between the time of fault occurrence and detection of faults in the FDI process. For the case of multiple faults, decision making based on only one inconsistency may cause problem. For example, if two faults, i.e., an internal leakage fault in R_4 and a sensor fault in β_{θ_2}, happen together in the steering system. The leakage fault is abrupt and the sensor fault is incipient. At time t_0, a coherence vector [0 1 0] is observed and from Table I, the internal leakage fault is detected and isolated. However, the sensor fault is also injected at time t_0; it cannot be detected at that time due to the hide effect of measurement noise. To achieve more reliable FDI, a waiting time T_{wt} is introduced as

$$Out = In(t_D + T_{wt}) \tag{29}$$

where t_D is the time point when a fault is detected.

The underlying meaning of the above equation is that the final fault detection result is based on the observation until time $t_D + T_{wt}$ instead of time t_D to accommodate the hide effect of measurement noise for incipient faults. Usually, $t_D = t_0$ if the system contains abrupt fault or intermittent fault. Finally, with the aid of waiting time, the residual is able to detect the sensor fault within the waiting time, and the new observed coherence vector [0 1 1] which includes β_{θ_2} as a fault candidate.

D. Prognosis Scheme

In this paper, model-based prognosis is used. Model-based methods use mathematical models to predict the fault growth trend. Given a proper model for a specific system, model-based methods can provide accurate prediction estimates. Compared with data-driven prognosis approaches which usually need costly run-to-failure data, model-based approaches can provide physical insight of fault degradation and hence can address subtle performance problems when a correct and accurate degradation model is available. However, in some cases, it is difficult or even impossible to build physical degradation models related to the monitored systems [18]. In this paper, some prescribed models in (16)–(19) are deployed for the behaviors of parameters and/or β evolving over time. Since in most cases, the degradation mechanism is difficult to obtain in advance, till now, only a limited number of degradation models exist, such as battery state of charge degradation and crack progression. When more degradation mechanisms are available, they can be fed back in the four dynamic behaviors to achieve more accurate prediction. For example, the battery degradation is usually modeled as an exponential growth process for the internal battery parameter which is typically inversely proportional to capacity [19]. This solution of the fault profile in (16) is also exponential form which is similar to the battery degradation process. This analogy is used to describe the trend of the degradation process in the vehicle steering system. In addition, the degradation process of different component/subsystem in the vehicle steering system may follow other form of dynamics; thus, some other possible dynamic equations (linear or nonlinear) are also used and the coefficients of these models are identified using collected data.

The purpose of prognosis is to predict the RUL of the faulty component based on the identified degradation model provided that a failure threshold is set [20]–[24]. This module is only activated when one or more FDV vectors from the AHDE estimator are [0 0 1] which indicates that at least one fault exhibits incipient nature from the identification result. RUL can be calculated by subtracting current time and data collection period from the time when the fault reaches the failure value. The objective of prognosis is to predict EOL at a certain time point t_f using the observation up to time $t_f + N \cdot t_s$, where N is the number of data collected and t_s is the sampling time. Let $T_{EOL} = 1$ denote the event that a failure threshold is exceeded, and 0 otherwise. EOL and RUL can be formulated as

$$EOL(t_f) = \inf\{t \in \mathbb{R} : t \geq (t_f + N \cdot t_s) \wedge T_{EOL} = 1\} \tag{30}$$

$$RUL(t_f) = EOL(t_f) - (t_f + N \cdot t_s). \tag{31}$$

Note that from (19), it is known that usually there is certain delay between the time of fault occurrence and detection of faults if the monitored system contains incipient faults. If the system contains abrupt fault or intermittent fault other than incipient fault, i.e., $t_D = t_0$, the data is preferred to be collected from time point t_0, i.e., $t_f = t_0$, to avoid the estimation of initial value for incipient fault. In other words, if the data are collected after $t_0 + T_{wt}$, the initial value for the incipient fault is not equal to its nominal one and also needs to be estimated. To help better understand (30) and (31), let us consider the internal leakage fault in the hydraulic cylinder of the steering system. The partial cylinder failure can be due to worn seal, define R_{4F} as the failure threshold for the fault and R_{4F} means the smallest value of R_4 at which the hydraulic cylinder operates within the functional limits. Thus, $T_{EOL} = 1$ if $R_4 < R_{4F}$. The fault in the hydraulic cylinder R_4 will degrade (slowly decrease for this case) according to a nonlinear form as $\dot{R_4} = b_2 R_4^2$ in (17). With the estimated degradation model and coefficients, the RUL is computed as $(1/R_4(0) - 1/R_{4F})/\bar{b}_2 - N \cdot t_s$, where $\bar{b}_2$ is the estimated value of b_2 and $R_4(0)$ is the value of R_4 at time point t_f.

IV. Experimental Results

The fault diagnosis and prognosis strategies given in Section III are experimentally tested. The physical parameters of the steering system are presented in Table III. In the experimental test bench, four sensors (one absolute encoder, one incremental encoder, and two pressure sensors) are used to collect different signals as shown in Fig. 1. A notebook with Window operating system hosts the online FDI software. The software uses MatlabR2007B with Simulink xPC Target to run and control real-time applications on the target PC mounted on platform of the steering system. The host PC, i.e., the notebook, builds the FDI module using Simulink functions. The Microsoft Visual C++ compiler in the host PC creates executable code according to the Simulink FDI module. The executable code is downloaded from the host PC to the target PC to run the xPC Target real-time kernel. After downloading the executable code, one can run and test the target application in a real-time manner. The target PC carries out the real-time residual computation

TABLE III
VALUES OF THE NOMINAL PHYSICAL PARAMETERS

Parameter	Value	Unit
kf_2	1.017e^{-5}	Nm/(rad/s)
Fu_2	6.20e^{-2}	Nm
kf_3	3.161e^{-1}	Nm/(rad/s)
Fu_3	3.55e^{-2}	Nm
kf_5	33.28	Nm/(rad/s)
Fu_5	22.11	Nm
J_1	3.727e^{-5}	kg·m^2
J_2	5.627e^{-4}	kg·m^2
J_3	9.78e^{-2}	kg·m^2
k_1	3	A/V
k_2	5.627e^{-4}	Nm/A
k_3	4/70	/
k_4	1/2	/
k_5	1.024e^{-4}	m^3
A	4.46e^{-4}	m^2

using measurements acquired from the sensors. All the measurements from the sensors are interfaced with the target PC via the PCI cards (NI PCI-6025E and NI PCI-6713). The host PC performs the online FDI decision making as well as fault estimation and prognosis using measurements acquired from the target PC. The host computer accesses the low-level target PC (on-board computer) through wireless communication using local wireless network. The steering system is an open-loop system. The motion of the dc motor of the steering system is generated by a current-modulated PWM voltage from a current amplifier. The command input signal is defined as follows (in Volts):

$$u_{in} = 0.17(5 \sin 2t + 0.5 \sin t + 0.2 \sin 0.5t + \sin 2.5t). \tag{32}$$

A single internal leakage fault with abrupt change is performed. The purpose of this experiment is not to validate the proposed methods, but to provide the fault value for R_4 under abrupt fault condition. This value will act as a reference for further experiment. In addition, the position of the valve under this fault condition is marked. Internal leakage is a common fault in hydraulic actuators. The piston of the hydraulic actuator divides the cylinder into two chambers which are filled with oil. When the pump rotates in one direction, the oil at the high pressure side pushes the piston which drives the steering wheels through a steering mechanism as shown Fig. 1. The oil at the low pressure side returns to the oil reservoir inside the pump through the low pressure tube. Under normal condition, oil is not supposed to flow from one chamber to the other. The piston efficiency is decreased if the internal flow between the chambers is present. To simulate the internal leakage fault in the steering system, an oil bypass has been installed to allow free oil flow from one side to the other side of the cylinder. A manual valve is used to control the oil flow in the bypass tube. Under normal condition, the valve is completely closed. An open valve represents abnormal condition, and the fault severity is determined by the opening level of the valve. Fig. 4 shows the response of residuals. The identified $R_{4\_ab} = 2.5221e^7$ kg$^{-(1/2)}$m$^{-(1/2)}$ as shown in Fig. 5. There are totally three fault scenarios considered in this study.

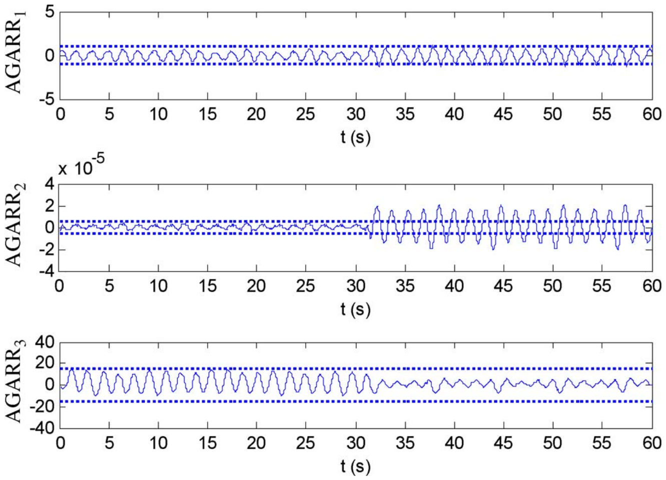

Fig. 4. Residual responses under an abrupt internal leakage fault.

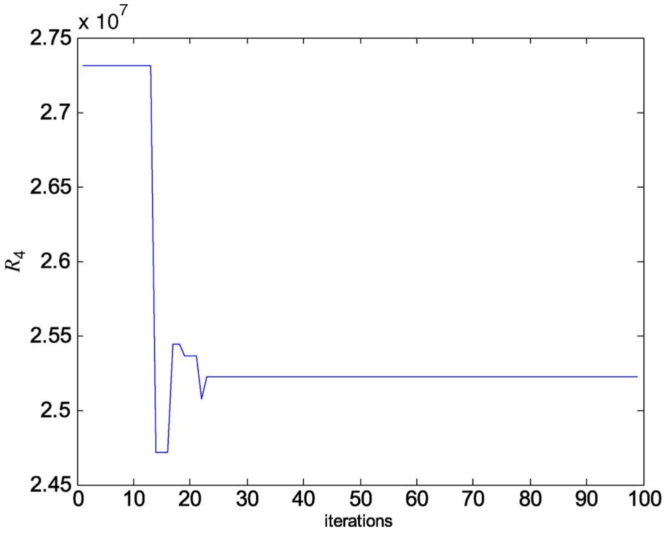

Fig. 5. Identification for R_4.

A. Scenario One: Abrupt Internal Leakage Fault and Intermittent Sensor Fault

For the first scenario, two faults of different type occur simultaneously in the steering system. These faults include an abrupt internal leakage fault of the piston and an intermittent sensor fault in $Df : \theta_1$. The leakage fault is introduced by turning the valve to the marked position mentioned previously, while the sensor fault is introduced by multiplying the sensor measurement by its efficiency factor β_{θ_1}. The fault profile of the intermittent sensor fault is demonstrated in Fig. 6. Both faults are introduced at $t = 32$ s and the knowledge of fault types is unavailable for the designer.

Fig. 7 presents the residual responses under the abrupt internal leakage fault and intermittent sensor fault in $Df : \theta_1$ where dashed lines indicate the thresholds. The thresholds are set as $\epsilon_1 = 1$, $\epsilon_2 = 5.5e^{-6}$ and $\epsilon_3 = 15$. Usually, the thresholds are set by observation of system responses under normal condition and they should be defined carefully to avoid false alarm. After 32s, a coherence vector [1 1 0] is observed. From Table II, this coherence vector may be caused by the broken belt fault. Thus, the fault mode estimator is activated and the

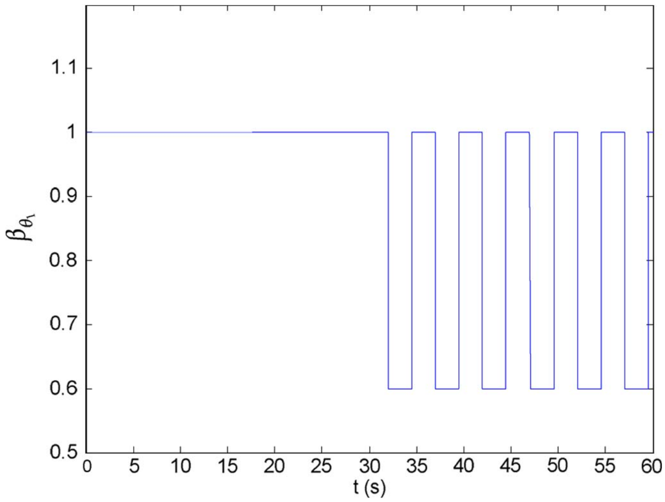

Fig. 6. Fault profile in sensor $Df:\theta_1$.

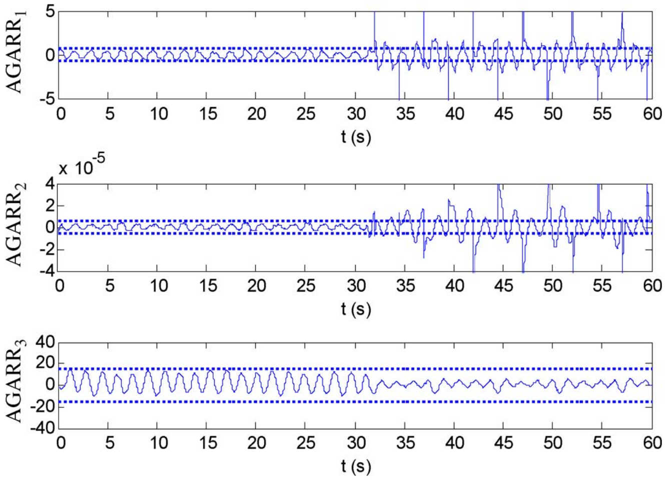

Fig. 7. Residual responses under an abrupt internal leakage fault and an intermittent sensor fault in $Df:\theta_1$.

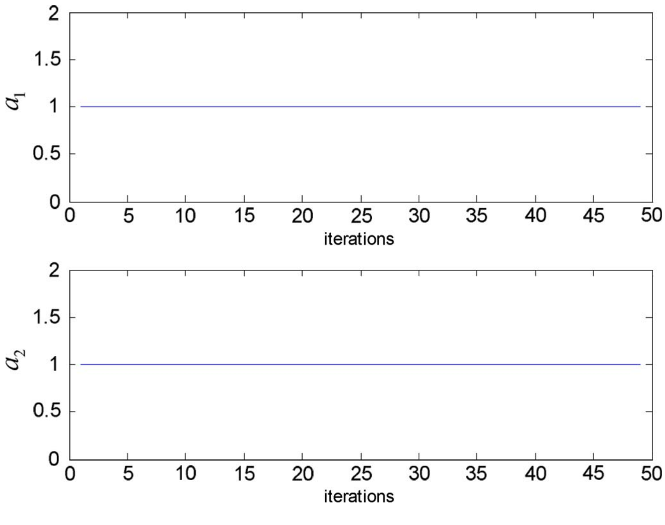

Fig. 8. Fault mode identification result.

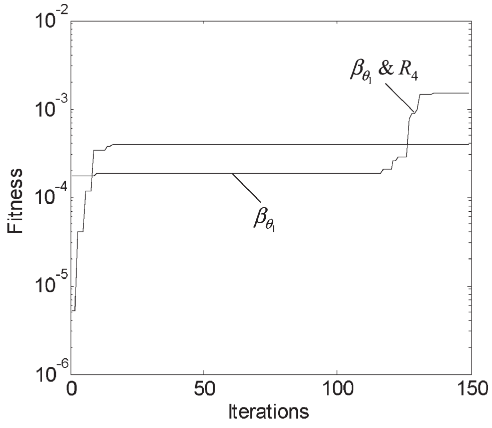

Fig. 9. Fitness evolution for the first fault scenario.

TABLE IV
EXPERIMENTAL RESULTS FOR THE FIRST SCENARIO

Fault in β_{θ_1}	Designed Value	Estimated Value
FDV	[0 1 0]	[0 1 0]
T	5s	5.27s
$\beta_{\theta_1\_int}$	0.6	0.623
Fault in R_4	Designed Value	Estimated Value
FDV	[1 0 0]	[1 0 0]
R_{4_ab}	$2.5221\text{e}^7\text{kg}^{-\frac{1}{2}}\text{m}^{-\frac{1}{2}}$	$2.6727\text{e}^7\text{kg}^{-\frac{1}{2}}\text{m}^{-\frac{1}{2}}$

identified mode is $[a_1\ \ a_2] = [1\ 1]$ as shown in Fig. 8. Since no new mode is identified, the inconsistency is caused by the faults at mode $[a_1\ \ a_2] = [1\ 1]$. For the steering system, the knowledge of the fault types is unavailable in advance, a waiting time $T_{wt} = 5$ s is set to accommodate the hide effect of measurement noise and the slowly developing nature of the incipient fault. At 37s, the detected signature is also [1 1 0], then the fault set could be $\delta = [\beta_{\theta_1}, \beta_{\theta_1}\&R_4]$. Any of the element in the fault set, i.e., β_{θ_1} or $\beta_{\theta_1}\&R_4$, may lead to observed inconsistency. Note that the spikes that are present in residuals AGARR_1 and AGARR_2 are due to the discontinuous behavior of the intermittent fault and the calculation of its derivatives. To identify the true faults, two AHDE estimators run in parallel and each estimator corresponds to one element in the fault set. One thousand sample data $(N = 1000)$ are collected with sampling time $t_s = 0.01$ s for the identification purpose. The parameters associated with AHDE are chosen as: Population size $= 100$, Maximum iterations $= 150$. The other control parameters in AHDE are dynamically adjusted by the adaptive scheme described in (25)–(27) according to the improvement measure defined in (28). The evolution of the fitness value versus iterations is illustrated in Fig. 9. From the figure, it is observed that the estimator for element $\beta_{\theta_1}\&R_4$ achieves higher fitness value compared with the one from the estimator for element β_{θ_1}. Therefore, the true faults are β_{θ_1} and R_4. The identified faults together with their FDVs are summarized in Table IV. In the table, the estimated period and the fault amplitude of the intermittent fault in β_{θ_1} are 5.27s and 0.623, which are close to their designed values. The abrupt internal leakage fault is also accurately estimated.

B. Scenario Two: Abrupt Internal Leakage Fault and Incipient Sensor Fault

In the second fault scenario, an incipient sensor fault in sensor $Df:\theta_2$ and an abrupt internal leakage fault in piston

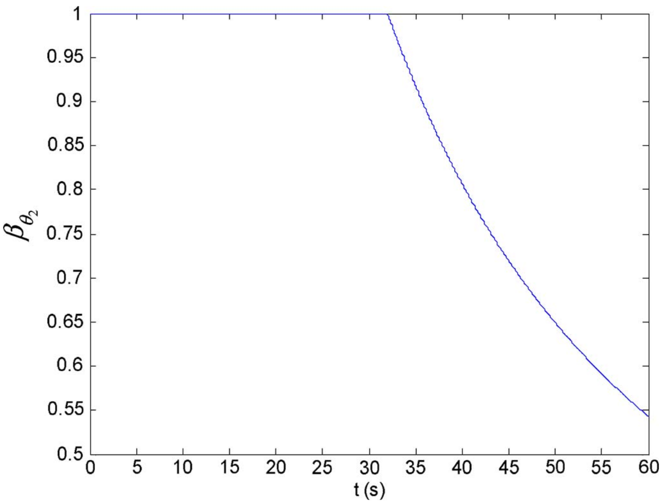

Fig. 10. Fault profile in sensor $Df : \theta_2$.

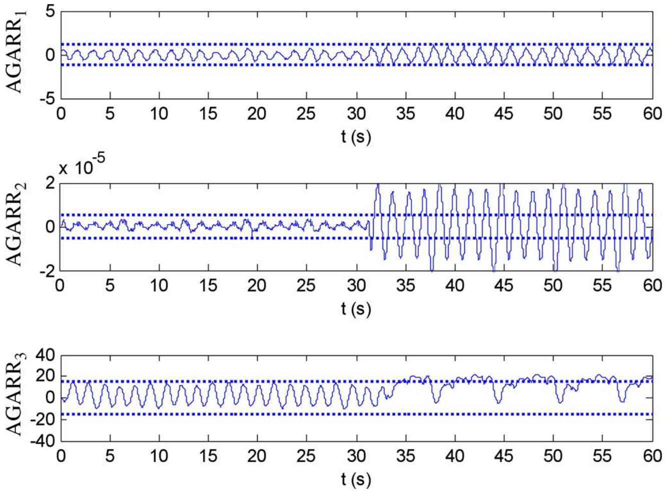

Fig. 11. Residual responses under an abrupt internal leakage fault and an incipient sensor fault in $Df : \theta_2$.

are considered. These two faults are injected at $t = 32$ s and the fault profile of sensor fault is presented in Fig. 10. The dynamics of the incipient sensor fault matches the nonlinear equation described in (17) with $b_2 = -0.03$. The failure value is chosen as $\beta_{\theta_2 F} = 0.5$. Fig. 11 demonstrates the residual responses under these two faults. Since the fault type for any fault is unknown, the information of the presence of incipient fault in the steering system is not available. Thus, a waiting time is required to accommodate the hide effect of measurement noise if the system contains incipient faults. The waiting time is set as $T_{wt} = 5$ s. Around $t = 32$ s, a coherence vector [0 1 0] is detected. The monitoring process is carried on within the waiting time, i.e., $t = 37$ s, and a new coherence vector [0 1 1] is observed around $t = 35$ s. From Table II, it is clear that this coherence vector [0 1 1] is not caused by a fault mode. As a result, the fault set could be $\delta = [\beta_{\theta_2}, \beta_{\theta_2}\&R_4, \beta_{\theta_2}\&R_5, \beta_{\theta_2}\&R_4\&R_5, R_4\&R_5]$ according to Table I. Since the system contains abrupt fault or intermittent fault other than incipient fault from the residual responses, the data are collected from $t = 32$ s to avoid the estimation of initial value for incipient fault. In the fault identification, 1000 sample data $(N = 1000)$ are collected and five AHDE estimators run in parallel, each based on one element in the fault set. The evolution of the fitness values for different AHDE estimators is shown in Fig. 12. It is concluded that the true faults are $\beta_{\theta_2}\&R_4$ because the estimator associated with this element achieves highest fitness value among all the estimators according to Fig. 12. The identified results are listed in Table V. From the table, the estimated vector for degradation dynamics of β_{θ_2} is $[\eta_1\ \eta_2] = [0\ 1]$ which matches the designed value. The estimated RUL is also close to the designed one.

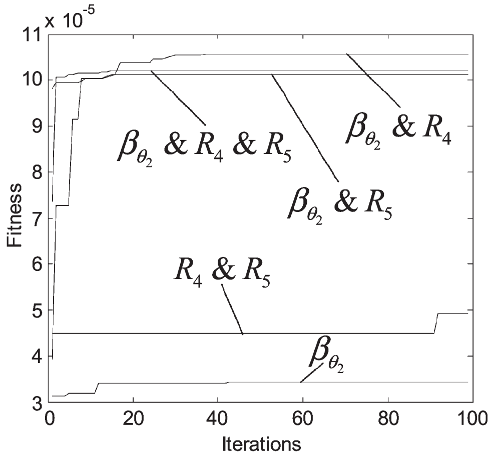

Fig. 12. Fitness evolution for the second fault scenario.

TABLE V
EXPERIMENTAL RESULTS FOR THE SECOND SCENARIO

Fault in β_{θ_2}	Designed Value	Estimated Value
FDV	[0 0 1]	[0 0 1]
$[\eta_1\ \eta_2]$	[0 1]	[0 1]
b_2	-0.03	-0.0286
RUL	23.34s	24.97s
Fault in R_4	Designed Value	Estimated Value
FDV	[1 0 0]	[1 0 0]
R_{4_ab}	$2.5221\mathrm{e}^7\mathrm{kg}^{-\frac{1}{2}}\mathrm{m}^{-\frac{1}{2}}$	$2.5637\mathrm{e}^7\mathrm{kg}^{-\frac{1}{2}}\mathrm{m}^{-\frac{1}{2}}$

C. Scenario Three: Burnt Driver Fault

A burnt driver fault is considered in the third fault scenario, in which the driver is suddenly disabled at $t = 32$ s. This fault causes that there is no power delivered to dc motor and the system is completely failed under this condition. Fig. 13 shows the residual responses under this fault mode. After $t = 32$ s, a coherence vector [1 0 0] is detected and from Tables I and II, it is clear that the signature is only caused by the burnt driver fault which is modeled as a fault mode. Thus, the burnt driver fault is detected and isolated.

D. Comparison Study

To demonstrate the performance of AHDE-based fault identification and prognosis method without the information of fault types, the AHDE has been compared with GA and AGA [17]. All the methods are evaluated on the same experimental data which are collected for the second fault scenario. Since the true faults are β_{θ_2} and R_4, then all the methods are carried

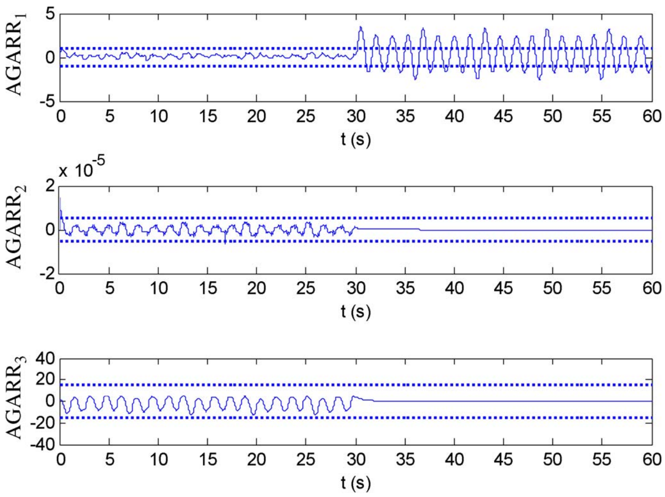

Fig. 13. Residual responses under a burnt driver fault.

out only for element $\beta_{\theta_2}\&R_4$. To apply GA, possible solutions are encoded in the form of strings, referred to as individuals, and the quality of solutions is measured by the fitness function defined in (24). The binary-number encoding method is used, in which the degradation model coefficients for incipient fault, fault values for abrupt fault and intermittent fault, and intermittent fault period are expressed in binary chromosomes consisting of 5 bits each, while a binary chromosome with 3 bits for FDV (each bit corresponds to one element in FDV) and a binary chromosome with 2 bits for η (each bit corresponds to one element in η). Note that since FDV and η are binary numbers, thus decoding is not required for the chromosomes related to FDV and η during the search process.

The construction of GA to cope with the identification problem consists of four major modules: initialization and encoding, evaluation, selection and reproduction, and crossover and mutation. In initialization and encoding, two sets of binary chromosomes (one for β_{θ_2} and the other for R_4), each set of binary chromosomes consists of all the parameters for encoding mentioned previously, form the individuals in the GA. During evaluation, each individual is evaluated by the fitness function defined in (24). For selection and reproduction, individuals with the highest fitness values are retained in the next generation, while those with the lowest fitness values are discarded. Here, the elitism mechanism is adopted, so the fitness of the current and previous best individuals are compared. If a fitter individual is not generated, the previous best individual is kept in the new population. As for the rest population, individuals are copied according to their fitness values The meaning of copying according to their fitness values is that individuals with a higher value have a higher probability of contributing one or more offspring in the next generation.

The well-performing individuals are granted a greater chance to recombine with other individuals, to reproduce offspring by using the genetic operators of crossover and mutation. The crossover operator performs combination. One-point crossover is used in this paper. Another way to cause individuals created during a reproduction to differ from their parents is mutation. Mutation is occasional (with small probability) random alteration of the value of an individual position. It is defined by the user and usually the lower the rate is, the less chance the individuals of the children differ from those of their parents. The structure of AGA is same with the structure of GA; the AGA can adaptively adjust the crossover and mutation rates based on the performance of the current genetic operators. It is able to increase the probability of the genetic operator if it consistently produces a better offspring during the search process. On the contrary, it attempts to decrease the probability of the genetic operator if it always produces a poorer offspring. This approach can adaptively adjust the balance between the exploration and exploitation of the solution space [17].

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT METHODS

	$F^{min}_{fitness}$	$F^{max}_{fitness}$	$F^{mean}_{fitness}$	σ
AHDE	10.33e^{-5}	10.72e^{-5}	10.48e^{-5}	1.265e^{-6}
GA	9.66e^{-5}	10.23e^{-5}	10.13e^{-5}	2.324e^{-6}
AGA	10.12e^{-5}	10.56e^{-5}	10.37e^{-5}	1.702e^{-6}

To carry out a fair comparison, the population size and the maximum iteration of all approaches are set equal to 150 and 300, respectively. The crossover probability $p_c = 0.55$ and mutation probability $p_m = 0.08$ in GA. For each algorithm, 20 independent runs are performed. Table VI summarizes the minimum fitness $F^{min}_{fitness}$, the maximum fitness $F^{max}_{fitness}$, and the average fitness $F^{mean}_{fitness}$ over the 20 runs. In the table, the standard deviation values σ of different approaches are also given. It is observed that AHDE outperforms the other algorithms in terms of final solution and standard deviation.

V. CONCLUSION

In this paper, a model-based multiple fault diagnosis and prognosis method is developed for the steering system of a CyCab electric vehicle. This method does not require the prior information of fault types and the degradation dynamics for the incipient faults. A new AHDE algorithm with less control parameters is proposed to find the true faults together with their types. Once a fault is detected, the AHDE estimator can estimate the fault magnitude for the abrupt fault, the period, and fault magnitude for the intermittent fault as well as the degradation dynamics for the incipient faults. After that, prognosis is carried out based on the identified degradation dynamics if the system has incipient faults. Three different fault scenarios have been experimentally tested to illustrate the effectiveness of the proposed method.

REFERENCES

[1] D. Brambilla, L. M. Capisani, A. Ferrara, and P. Pisu, "Fault detection for robot manipulators via second-order sliding modes," *IEEE Trans. Ind. Electron.*, vol. 55, no. 11, pp. 3654–3693, Nov. 2008.

[2] S. Arogeti, D. Wang, C. B. Low, and M. Yu, "Fault detection isolation and estimation in a vehicle steering system," *IEEE Trans. Ind. Electron.*, vol. 59, no. 12, pp. 4810–4820, Dec. 2012.

[3] C. B. Low, D. Wang, S. Arogeti, and J. B. Zhang, "Causality assignment and model approximation for hybrid bond graph: Fault diagnosis perspectives," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 3, pp. 570–580, Jul. 2010.

[4] C. B. Low, D. Wang, S. Arogeti, and M. Luo, "Quantitative hybrid bond graph-based fault detection and isolation," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 3, pp. 558–569, Jul. 2010.

[5] A. K. Samantaray and B. Ould Bouamama, *Model-Based Process Supervision: A Bond Graph Approach*. London, U.K.: Springer-Verlag, 2008.

[6] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, Dec. 1997.

[7] L. Zhang, Y. C. Jiao, Z. B. Weng, and F. S. Zhang, "Design of planar thinned arrays using a Boolean differential evolution algorithm," *IET Microw., Antennas Propag.*, vol. 4, no. 12, pp. 2172–2178, Dec. 2010.

[8] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15, no. 1, pp. 4–31, Feb. 2011.

[9] O. F. Eker, F. Camci, A. Guclu, H. Yilboga, M. Sevkli, and S. Baskan, "A simple state-based prognostic model for railway turnout systems," *IEEE Trans. Ind. Electron.*, vol. 58, no. 5, pp. 1718–1726, Mar. 2011.

[10] M. A. Djeziri, R. Merzouki, and B. Ould-Bouamama, "Robust monitoring of an electric vehicle with structured and unstructured uncertainties," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4710–4719, Nov. 2009.

[11] J. H. Luo, K. R. Pattipati, L. Qiao, and S. Chigusa, "Model-based prognostic techniques applied to a suspension system," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 5, pp. 1156–1168, Sep. 2008.

[12] M. Yu, D. Wang, M. Luo, and L. Huang, "Prognosis of hybrid systems with multiple incipient faults: Augmented global analytical redundancy relations approach," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 3, pp. 540–551, May 2011.

[13] B. Zhang, T. Taimoor, R. Patrick, G. Vachtsevanos, M. Orchard, and A. Saxena, "Application of blind deconvolution de-noising in failure prognosis," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 2, pp. 303–310, Feb. 2009.

[14] S. Jiang, R. Kumar, and H. E. Garcia, "Diagnosis of repeated/intermittent failures in discrete event systems," *IEEE Trans. Robot. Autom.*, vol. 19, no. 2, pp. 310–323, Apr. 2003.

[15] R. Merzouki, M. A. Djeziri, and B. Ould-Bouamama, "Intelligent monitoring of electric vehicle," in *Proc. IEEE/ASME Int. Conf. AIM*, 2009, pp. 797–804.

[16] P. J. Mosterman and G. Biswas, "Behavior generation using model switching: A hybrid bond graph modeling technique," *Trans. Soc. Simul.*, vol. 27, no. 1, pp. 177–182, Apr. 1995.

[17] K. L. Mak, Y. S. Wong, and X. X. Wang, "An adaptive genetic algorithm for manufacturing cell formation," *Int. J. Adv. Manuf. Technol.*, vol. 16, no. 7, pp. 491–497, Jun. 2000.

[18] A. K. S. Jardine, D. Lin, and D. Banjevic, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mech. Syst. Signal Process.*, vol. 23, no. 3, pp. 724–739, Apr. 2009.

[19] J. Zhang and J. Lee, "A review on prognostics and health monitoring of Li-ion battery," *J. Power Sources*, vol. 196, no. 15, pp. 6007–6014, Aug. 2011.

[20] M. Yu, D. Wang, and M. Luo, "Model based prognosis for hybrid systems with mode-dependent degradation behaviors," *IEEE Trans. Ind. Electron.*, vol. 61, no. 1, pp. 546–554, Jan. 2014.

[21] D. Wang, M. Yu, C. B. Low, and S. Arogeti, *Model-based Health Monitoring of Hybrid Systems*. New York, NY, USA: Springer-Verlag, 2013.

[22] M. Orchard, P. Hevia-Koch, B. Zhang, and L. Tang, "Risk measures for particle-filtering-based state-of-charge prognosis in lithium-ion batteries," *IEEE Trans. Ind. Electron.*, vol. 60, no. 11, pp. 5260–5269, Nov. 2013.

[23] C. Chen, B. Zhang, G. Vachtsevanos, and M. Orchard, "Machine condition prediction based on adaptive neuro-fuzzy and high-order particle filtering," *IEEE Trans. Ind. Electron.*, vol. 58, no. 9, pp. 4353–4364, Sep. 2011.

[24] B. Zhang, C. Sconyers, C. Byington, R. Patrick, M. Orchard, and G. Vachtsevanos, "A probabilistic fault detection approach: Application to bearing fault detection," *IEEE Trans. Ind. Electron.*, vol. 58, no. 5, pp. 2011–2018, May 2011.

Ming Yu (M'12) received the B.E. and M.S. degrees from Hefei University of Technology, Hefei, China, in 2001 and 2004, respectively. He received the Ph.D. degree from Nanyang Technological University, Singapore, in 2012.

Currently, he is a research fellow at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests are in fault diagnosis and prognosis, nonlinear control, and evolutionary algorithms.

Danwei Wang (S'88–M'89–SM'04) received the B.E degree from the South China University of Technology, Guangzhou, China in 1982, and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor in 1984 and 1989, respectively.

Currently, he is the Deputy Director of the Robotics Research Center, NTU, the Deputy Director for the EXQUISITUS, the Centre for E-City, and the Program Director for E-Mobility. His research interests include fault diagnosis and prognosis, robotics, control theory, and applications.

Prof. Wang is an Associate Editor of the *International Journal of Humanoid Robotics*. He has served as General Chairman, Technical Chairman, and various positions in the organizing committees of several international conferences, such as International Conference on Control, Automation, Robotics, and Vision, and IEEE International conference on Robotics and Automation. He was a recipient of the Alexander von Humboldt Fellowship, Germany.