RESEARCH PAPER

T. Zhang · K. K. Choi · S. Rahman · K. Cho · P. Baker · M. Shakil · D. Heitkamp

A hybrid surrogate and pattern search optimization method and application to microelectronics

Abstract A hybrid method for robust and efficient optimization process is developed by integrating a new response surface method and pattern search algorithm. The method is based on: (1) multipoint approximations of the objective and constraint functions, (2) a multiquadric radial basis function (RBF) for the zeroth-order function approximation and a new RBF plus polynomial-based moving least-squares approximation for the first-order enhanced function approximation, and (3) a pattern search algorithm to impose a descent condition and applied adaptive subregion management strategy. Several numerical examples are presented to illustrate accuracy and computational efficiency of the proposed method for both function approximation and design optimization. To demonstrate the effectiveness of the proposed hybrid method, it is applied to obtain optimum designs of a microelectronic packaging system. A two-stage optimization approach is proposed for the design optimization. The material properties of microelectronic packaging system and the shape parameters of solder ball are selected as design variables. Through design optimization, significant improvements of durability performances are obtained using the proposed hybrid optimization method.

Keywords Radial basis function · Moving least-squares method · Pattern search · Hybrid optimization method · Design variables · Microelectronics

Nomenclature

$\mathbf{x}$	A vector of design variables
$F_0(\mathbf{x})$	Objective function
$F_j(\mathbf{x})$	jth constraint function
A_i	Lower bound of the ith design variable
B_i	Upper bound of the ith design variable

T. Zhang · K. K. Choi (✉) · S. Rahman
Department of Mechanical & Industrial Engineering,
The University of Iowa, Iowa City, IA 52242, USA
Tel.: +1-319-3355684
e-mail: kkchoi@engineering.uiowa.edu

K. Cho · P. Baker · M. Shakil · D. Heitkamp
Government Systems Division, Rockwell Collins, Inc., Cedar Rapids,
IA 52498, USA

$\tilde{F}_0^k$	Approximation of the original objective function at iteration k
$\tilde{F}_j^k$	Approximation of the jth constraint function at iteration k
$A_i^{(k)}$	Lower bound of the subregion at iteration k
$B_i^{(k)}$	Upper bound of the subregion at iteration k
$\mathbf{x}_*^{(k)}$	Optimum solution at iteration k
$r_j^{(k)}$	Error measure for jth function at iteration k
ϕ	Radial basis function
λ_j	Associated radial basis coefficient corresponding to the jth data point
$\tilde{u}(\mathbf{x})$	Approximation by moving least-squares method
$p_i(\mathbf{x})$	Basis function of moving least-squares method
$b_i(\mathbf{x})$	Coefficient for the basis function of moving least-squares method
ε	Error measurement for the function approximation
$m(\boldsymbol{u})$	Total mass of the torque arm
Δu_i	Global relative displacements in x, y, and z directions
α	Coefficients of thermal expansion of selected materials
$\boldsymbol{D}$	Shape parameters of solder ball
N_f	Fatigue crack-initiation life of the electronic packaging system

1 Introduction

Modern design, engineering analysis, and optimization problems involve computationally expensive physics-based simulation, in which the simulation of one design can consume a significant amount of computational time and require a large amount of human effort. Thus, efficient and systematic methods for performing design studies to find an optimum design or select the best compromise between competing objectives can yield significant improvement in design cycles. For design optimization, design sensitivity analysis (Choi and Kim 2004) is an important and costly procedure in the

conventional gradient-based algorithm. Recently, significant progress has been made in linear and nonlinear sensitivity analysis of engineering structures (Choi and Kim 2002; Kim et al. 2000, 2001). However, the sensitivity analysis of non-linear and path-dependent structural systems is still a very much challenging problem. To address this problem, alternative methods such as response surface methodology (RSM) (Myers and Montgomery 1995) and direct search methods (Lewis et al. 1998; Box 1957) have been developed.

RSM approximates the relationship between a response (dependent variables) and a vector of input variables (independent variables) by constructing smooth functions of objective and constraints using limited sample points. RSM is effective when the function evaluation is computationally expensive and the calculation of the design sensitivity information is difficult. As a result of necessity, there is growing interest in using response surface techniques in design optimization. Two major trends have emerged in the use of RSM: (1) the use for global approximations (Giunta et al. 1994; Liu et al. 2004; Venter et al. 1998), where response surfaces are constructed on the entire design space; and (2) the use for local approximations (Toropov et al, 1993; Youn and Choi 2004), where response surfaces are built on a local subregion around the current design. The cost of developing an accurate global response surface function is higher than that for local responses because it is more expensive to sample in the whole design space for the global accuracy of non-linear response. Response surface functions are usually only accurate over a part of the design space called the subregion. Hence, local sequential response surface models need to be developed for multipoint approximations. The objective and constraint functions in each iteration are approximated by response surface functions. The solution of the approximated optimization problem becomes the starting point for the next step, until an optimum design is reached. The response surface approximations are updated with new information in the neighborhood of the current design. A primary concern in developing such an approximation optimization procedure is the proper choice of the subregion management strategy. Indeed, inappropriate strategies can result in premature convergence, often leading to inaccurate solutions. An important consideration is the strategy used to find an effective size and location of this region.

The direct methods, developed since the 1950s (Lewis et al. 1998; Box 1957), are also known as zeroth-order methods because no function derivatives are used. Pattern search method is a class of direct search methods. Since the introduction of the pattern search method, the method has remained popular with users. Recently, the fact that they are proven to be convergent has generated renewed interest in the nonlinear programming community (Lewis and Torczon 2000). However, all these advantages do not come without a price. Because the pattern search method is a discrete search algorithm, they do not entail rapid convergence properties (Frank et al. 1992). Furthermore, because pattern search can be viewed as sampling methods, they are less efficient for large-dimensional problems (Glowinski and Kearsley 1995).

In addition, it is difficult to find the search direction for general nonlinear constrained optimization problems (Lewis and Torczon 2000).

In this paper, a new hybrid method is presented by integrating the merits of RSM and pattern search methods for design optimization. The method is based on multipoint approximations of the objective and constraint functions, a multiquadric radial basis function (RBF) for zeroth-order function approximation or RBF plus polynomial-based moving least-squares (MLS) approximation for first-order function approximation, and a pattern search algorithm to impose a descent condition and applied adaptive subregion management strategy. The accuracy and efficiency of the proposed method are illustrated by several numerical examples.

Electronic assemblies are used in many applications with increased reliability expectations in more severe operating environments. The effect of temperature cyclic load on the reliability of electronic packaging has been the subject of much work (Lau 1991; Ohring 1998; Lau 1995). Under temperature cycles, a mismatch of the coefficients of thermal expansion (CTE) in the assembly can induce repeated stresses, resulting in fatigue damage accumulation in the solder joints. Progressive damage in solder joints eventually leads to device failure. Due to the increasingly competitive market pressures, more reliable and robust microelectronic designs are desirable. To resolve this issue, optimization of the whole assembly of microelectronic packaging systems becomes necessary. The proposed hybrid method is employed for design optimization of ASAT 144 fine-pitch ball-grid array (fpBGA) electronic packaging. The fatigue life is selected as the objective function, and the material properties of the printed circuit board (PCB) and package and the shape parameters of the solder joint are selected as the design variables. Different combinations of material properties and shape parameters are evaluated in several case studies. Through design optimization, the durability performance of the microelectronic packaging system is improved significantly.

2 A hybrid optimization method

In this paper, a hybrid method is proposed for solving the general constrained optimization problem, which is to

minimize $\quad F_0(\mathbf{x}), \ \mathbf{x} \in \Re^n$

subject to $\quad F_j(\mathbf{x}) \le 0, \ j = 1, \cdots, l$

and $\qquad A_i \le x_i \le B_i, \ i = 1, \cdots, n$ $\qquad$ (1)

where $\mathbf{x}$ is a vector of n design variables, $F_0(\mathbf{x})$ is the objective function, $F_j(\mathbf{x}), \ j = 1, \cdots, l$ are the constraint functions, and A_i and B_i are lower and upper bounds, respectively, of the ith design variable x_i. The hybrid optimization method involves: (1) RBFs for approximating objective and constraints functions in a subregion, (2) subregion movement obtained by both of RSM and pattern search, and (3) pattern search

algorithms for imposing descent condition explicitly and applied adaptive subregion management strategy.

2.1 Response surface-based design optimization

Using multipoint approximations (Toropov et al. 1993), the problem in (1) can be approximated to

minimize $\quad \tilde{F}_0^k(\mathbf{x}), \ \mathbf{x} \in \Re^n$

subject to $\quad \tilde{F}_j^k(\mathbf{x}) \leq 0, \ j = 1, \cdots, l$

and $\qquad A_i^{(k)} \leq x_i \leq B_i^{(k)}, \ i = 1, \cdots, n$

$$A_i^{(k)} \geq A_i; \ B_i^{(k)} \leq B_i, \ i = 1, \cdots, n \qquad (2)$$

where k is the current iteration number, $\tilde{F}_0^k(\mathbf{x})$ and $\tilde{F}_j^k(\mathbf{x})$ are explicit local response surface approximations of the original objective and constraints functions, respectively, at iteration k, and $A_i^{(k)}$ and $B_i^{(k)}$ are lower and upper bounds, respectively, of the subregion at iteration k. (2) can be solved using conventional gradient-based method. Let $\mathbf{x}_*^{(k)}$ denote the solution of (2) at iteration k. To evaluate the adequacy of response surface approximations, corresponding error measures are defined as

$$r_j^{(k)} = \left| \frac{F_j\left(\mathbf{x}_*^{(k)}\right) - \tilde{F}_j^{(k)}\left(\mathbf{x}_*^{(k)}\right)}{F_j\left(\mathbf{x}_*^{(k)}\right)} \right|, \ j = 0, 1, \cdots, l \qquad (3)$$

The next $(k+1)$th iteration is started from $\mathbf{x}_*^{(k)}$. However, the size and location of the next subregion, i.e., the move limits $A_i^{(k+1)}$ and $B_i^{(k+1)}$ at the $k+1$th iteration, depend on error estimates at the kth iteration, described as follows. Let

$$r_j^{(k)} \leq \varepsilon_j, \ j = 0, 1, \cdots, l \qquad (4)$$

define conditions of satisfactory response approximations at $\mathbf{x}_*^{(k)}$, where ε_j, $j = 0, 1, \cdots, m$ are given small positive numbers. If (4) does not hold even for any of the active constraints or the objective function, then the size of the search subregion at the $(k+1)$th iteration is reduced. When (4) is satisfied for all constraints and the objective function, the size of search subregion needs to be decided based on the location. If $\mathbf{x}_*^{(k)}$ is located inside the kth subregion (i.e., none of the move limits is active), the size of the next subregion is reduced. Otherwise, the search subregion must be moved in the direction $\mathbf{x}_*^{(k)} - \mathbf{x}_*^{(k-1)}$ with keeping its size or enlarging the size according to the accuracy. The iteration is terminated when: (1) (4) is satisfied for all $j = 0, 1, \cdots, l$, (2) move limits of the subregion are not active, and (3) the subregion has reached a required size, i.e.,

$$\max \left[\frac{B_i^k - A_i^k}{B_i - A_i} \right] \leq \delta, \ i = 1, \cdots, n \qquad (5)$$

where δ is a given small positive number.

The response surface model used in multipoint approximation provides a more accurate or efficient means of solving optimization problems than global response surface models. However, there are several important aspects, which may influence the accuracy and convergence rate of local response surface methods. They involve: (1) the choice of function approximation methods and (2) the move limit strategy. First, the approximate response functions need to be accurate. It should be noted that the accuracy requirements in local approximations can be achieved more easily than global ones, and obviously, more accurate function approximations would yield a faster convergence. Second, current response surface models do not incorporate descent conditions explicitly. This lack of descent condition can lead to unnecessary iterations or even to inaccurate solutions.

2.1.1 Zeroth-order function approximation

The RBF interpolation is a well-known method for function approximation of multivariate scattered data (Franke 1982). RBF provides interpolants to function values given at irregularly positioned data points and can be applied to any dimension. This makes RBF attractive in the theory of computational mathematics. Several types of RBFs are developed in the current literature. Among all methods tested, the multiquadric method ranks highest in accuracy (Franke 1982) and for this reason generated a great deal of interest among mathematicians (Light 1992; Wang and Liu 2002). Wang (2004) investigated the influence of shape parameter to the accuracy of function interpolation. Many researches are still under way.

The RBF interpolation is based on forming linear combinations of radial functions centered at each data point. For scattered data $F_j^{(i)} = F_j(x_i)$, which is available at $\mathbf{x}_i \in \Re^n$ for $i = 1, \cdots, N$ ($j = 0, ..., l$ are the objective and constraint functions in (1); $\mathbf{x}_i$ is a design point, i is the index of design points, n is the dimension of the design space, while N is the number of design points), RBF employs a "radial" function $\phi : \Re^n \to \Re$ to construct the interpolant

$$\tilde{F}_j(\mathbf{x}) = \sum_{i=1}^N \lambda_i \phi(\|\mathbf{x} - \mathbf{x}_i\|) \qquad (6)$$

where $\|\cdot\|$ is the Euclidean norm, $\phi(\|\mathbf{x} - \mathbf{x}_i\|)$ is a suitably chosen RBF centered at $\mathbf{x}_i$, and $\lambda_i \in \Re$ is the associated radial basis coefficient corresponding to the ith data point. Requiring $\tilde{F}_j(\mathbf{x}_i) = F_j^{(i)}$, the coefficients can be obtained by solving a system of linear equations (Light 1992).

In optimization problems, each RBF can be used to create response surface $\tilde{F}_j(\mathbf{x})$ that interpolates the true response function $F_j(\mathbf{x})$. The construction of such a response guarantees that the function value of $\tilde{F}_j(\mathbf{x})$ is identical to the values of both objective and constraint functions at the sample points. Alternative RBFs based on compact supports have been developed (Wendland 1995). The "compact support" is a term related to the weight function, which is nonzero only over a small subdomain around the sample point. To date, compactly supported positive-definite RBFs are available only for low-dimensional problems ($n \leq 3$), which is a

major limitation for realistic structural optimization problems. Moreover, the accuracy of the compactly supported RBF is not very good when compared with the results of globally supported RBF (Zhang et al. 2000). Indeed, development of compactly supported positive-definite RBFs in a high dimension is still an active research topic in the mathematical community. Nonetheless, RBF is an attractive choice for response surface approximation.

2.1.2 First-order enhanced function approximation

In some cases, sensitivity information can be obtained efficiently, and it can be used to improve the constructed response function. A polynomial-based response surface was proposed by van Keulen (Van Keulen et al. 2000). Chung and Alonso (2002) used gradients to construct the approximation by Kriging models. Liu and Batill (2000) developed a gradient-enhanced neural network response surface approximation. For the conventional polynomial bases, the basis selection is user dependent, while a large number of training points (sample points) are needed if the gradient-enhanced neural network is applied; for the gradient-based approximation by Kriging model, the most suitable design of experiments is still the ongoing research. Because of the good performance of RBF in zeroth-order function approximation, it is desirable to construct a first-order enhanced response surface function by using RBF. One inherent feature of RBF is that the number of total bases is equal to the number of sample points; each sample point is associated with an RBF. For the

interpolation of a zeroth-order, a given number of N sample points will result in N function values. Hence, N bases appear in the function approximation. In other words, the number of unknown coefficients is equal to the number of zeroth-order function values. However, if both function values and corresponding sensitivities are given at N sample points for an n-dimensional problem, then the total amount of information available is $N(1 + n)$, much greater than the information of function values. Because the conventional RBF is for zeroth-order function interpolation and the basis number is N, interpolation causes overdetermined system equations, a situation which has no solution. Therefore, it is necessary to use a least-squares-type technique to account for combined function and sensitivity information at all sample points. With the success of moving least squares (MLS) (Choi and Kim 2002; Kim et al. 2000) in computational mechanics and function approximation (Youn and Choi 2004), the MLS approximation is used in this paper.

Given N as the number of sample points with n as the number of design variables, there is $N(1 + n)$ amount of known information available. An approximation $\tilde{F}(\mathbf{x})$ of the function $F(\mathbf{x})$ in a domain Ω is defined as

$$\tilde{F}(\mathbf{x}) = \mathbf{p}^T(\mathbf{x})\mathbf{b}(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x})b_i(\mathbf{x}) \tag{7}$$

In (7), $p_i(\mathbf{x})$ is the basis, m is the number of bases, and $b_i(\mathbf{x})$ is the corresponding coefficient. To best fit $F(\mathbf{x})$, the coefficient vector $\mathbf{b}(\mathbf{x})$ is selected as the one that minimizes the error

$$J(\mathbf{b}(\mathbf{x})) = \sum_{I=1}^{N} \left\{ w_I^{(0)}(\mathbf{x}) \left[\mathbf{p}^T(\mathbf{x}_I)\mathbf{b}(\mathbf{x}) - F_I \right]^2 + w_I^{(1)}(\mathbf{x}) \left[\frac{d\mathbf{p}^T(\mathbf{x}_I)}{dx_1}\mathbf{b}(\mathbf{x}) - \frac{dF_I}{dx_1} \right]^2 + \cdots + w_I^{(n)}(\mathbf{x}) \left[\frac{d\mathbf{p}^T(\mathbf{x}_I)}{dx_n}\mathbf{b}(\mathbf{x}) - \frac{dF_I}{dx_n} \right]^2 \right\}$$

$$= \left[\mathbf{Q}\mathbf{b}(\mathbf{x}) - \hat{\mathbf{F}} \right]^T \mathbf{W}(\mathbf{x}) \left[\mathbf{Q}\mathbf{b}(\mathbf{x}) - \hat{\mathbf{F}} \right] \tag{8}$$

where

$$\mathbf{Q} = \begin{bmatrix} \mathbf{P} \\ \mathbf{P}_{x_1} \\ \vdots \\ \mathbf{P}_{x_n} \end{bmatrix}_{M \times m} \tag{9}$$

and $\mathbf{P}$ is the basis matrix with dimension $N \times m$, $\mathbf{P}_{x_i}$ is the matrix with dimension $N \times m$ such that each entry is the derivative of each entry of $\mathbf{P}$ with respect to design variable x_i, $\mathbf{W}(\mathbf{x})$ is an $M \times M$ matrix (cubic spline weight function is used in this paper), and each entry is the weight function with compact support (influence domain, which is nonzero only over a small subdomain around the sample point). The cubic spline weight function is defined as

$$w(r) = \begin{cases} \dfrac{2}{3} - 4r^2 + 4^3, \ r \leq \dfrac{1}{2} \\ \dfrac{4}{3} - 4r + 4r^2 - \dfrac{4}{3}r^3, \ r < \dfrac{1}{2} \leq 1 \\ 0, \ r > 1 \end{cases} \tag{10}$$

Define $d_I = \|\mathbf{x} - \mathbf{x}_I\|$, $r = d_I/d_{mI}$, where d_{mI} is the size of the domain of influence of the I-th point. The row number M is equal to $N(1 + n)$. In (8), $\hat{\mathbf{F}}$ is defined as

$$\hat{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{F}_{x_1} \\ \vdots \\ \mathbf{F}_{x_n} \end{bmatrix}_{M \times 1} \tag{11}$$

where $\mathbf{F}$ is the column vector of function values with dimension N, and $\mathbf{F}_{x_i}$ is the sensitivity vector with respect to design variable x_i. By applying the stationary condition to the weighted discrete error norm, the coefficient vector can be obtained by solving

$$\mathbf{A}(\mathbf{x})\mathbf{b}(\mathbf{x}) = \mathbf{B}(\mathbf{x})\hat{\mathbf{F}} \tag{12}$$

which can be inverted to obtain the following explicit form

$$\mathbf{b}(\mathbf{x}) = \mathbf{A}^{-1}(\mathbf{x})\mathbf{B}(\mathbf{x})\hat{\mathbf{F}} \tag{13}$$

where $\mathbf{b}(\mathbf{x})$ is an m-dimensional column vector and $\hat{\mathbf{F}}$ is an M-dimensional column vector, and A and B are the $m \times m$ and $m \times M$ matrices, respectively, defined by

$$\mathbf{A}(\mathbf{x}) = \mathbf{Q}^{\mathbf{T}}\mathbf{W}(\mathbf{x})\mathbf{Q} \qquad (14)$$

$$\mathbf{B}(\mathbf{x}) = \mathbf{Q}^{\mathbf{T}}\mathbf{W}(\mathbf{x}) \qquad (15)$$

In general, the accuracy of the approximation is decided by how the bases are selected and how many terms of the bases are selected. In the case of one-design variable, the basis definition is straightforward. However, if there is more than one-design variable, the interaction terms will arise. In a two-design variable case, bilinear term $x_1 x_2$ has to be included, while both bilinear and biquadratic terms need to be included for cubic bases. For example, the set of cubic bases is

$$P = \left\{ \begin{array}{l} 1, x_1, x_2, \cdots, x_n, x_1 x_2, x_1 x_3, \cdots, x_1 x_n, x_2 x_3, \cdots, x_2 x_n, \cdots x_{n-1} x_n, x_1^2, x_2^2, \cdots, \\ x_2^3, x_1 x_2 x_3, x_1 x_2 x_4, \cdots, x_1 x_2 x_n, x_1 x_3 x_4, \cdots, x_1 x_3 x_n, \cdots, x_1 x_{n-1} x_n, \\ x_2 x_3 x_4, \cdots, x_2 x_3 x_n, \cdots, x_2 x_{n-1} x_n, \cdots, x_{n-2} x_{n-1} x_n, x_1^2 x_2, x_1^2 x_3, \cdots, x_1^2 x_n, \\ x_2^2 x_1, x_2^2 x_3, \cdots, x_2^2 x_n, \cdots, x_{n-1}^2 x_1, x_{n-1}^2 x_2, \cdots, x_{n-1}^2 x_n, x_n^2 x_1, x_n^2 x_2, \cdots, x_n^2 x_{n-1}, x_1^3, x_2^3, \cdots, x_n^3 \end{array} \right\} \qquad (16)$$

The number of constant and linear terms is $(n+1)$, the number of bilinear terms is $(n!/1)n/2$, and the number of quadratic terms is n. In addition, there are $(n!/2)(n!/1)/2+(n!/3)(n!/2)/2+...+[n!/(n!/2)][n!/(n!/1)]/2$ triple linear terms and $n(n!/1)$ biquadratic terms, and the number of cubic terms is n.

The fundamental rule of bases selection is that the number of bases terms should not be larger than N for the case of zeroth-order and $N(1+n)$ for a first-order case. Moreover, the system equation, (12), must maintain nonsingularity. Due to the large number in polynomial basis combination discussed above, no clear approach yet exists for how to efficiently select the bases and maintain the accuracy of the approximation. In general, the selection of the basis is problem dependent and based on the users' experience.

To overcome the difficulty in bases selection, new types of bases are proposed in this paper that combines the multiquadric radial bases functions and the polynomial bases. The quadratic terms and the multiquadric RBF are not linearly independent and thus the system equation, (12), cannot maintain nonsingularity. Therefore, the proposed bases include constant, linear terms, bilinear terms, and cubic terms. The rest of the interaction terms in conventional polynomial basis are replaced by the RBF. These types of bases overcome the heuristics and uncertainty in selecting the bases. Using a systematic way for choosing bases, RBF can be applied in first-order function approximation. The generalized bases vectors in an n-dimension problem is defined as

$$\mathbf{p}^T(\mathbf{x}) = \left[1, x_1, \cdots, x_n, x_1 x_2, \cdots, x_1 x_n, x_2 x_3, \cdots, x_2 x_n, \cdots, x_{n-1} x_n, x_1^3, \cdots, x_n^3, p_1(\mathbf{x}), p_2(\mathbf{x}), \cdots, p_N(\mathbf{x}) \right] \qquad (17)$$

where

$$p_i(\mathbf{x}) = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2 + c^2} \qquad (18)$$

From (17), it is seen that the proposed bases are compact and systematic. The accuracy of the proposed bases will be demonstrated in forthcoming numerical examples.

2.2 Pattern search method

The pattern search method (Lewis et al. 1998) is characterized by a series of exploratory moves that consider the behavior of the objective function in a pattern of points, all of which lie on a rational lattice. For simplicity, first consider an unconstrained minimization problem, given by

$$\text{Minimize} \quad F_0(\mathbf{x}), \quad \mathbf{x} \in \Re^n \qquad (19)$$

where $F_0 : \Re^n \rightarrow \Re$ is a continuous and differentiable function, but the sensitivity (gradient) of F_0 is either unavailable or unreliable, and $\mathbf{x} = [x_1, \cdots, x_n]^T \in \Re^n$ is an n-dimensional real vector of the design variables. The pattern search algorithm involves three steps to solve (19). First, at iteration k, start with a point $\mathbf{x}_k \in \Re^n$ and a corresponding step-length parameter $\Delta_k > 0$. Second, if $\boldsymbol{e}_i, i = 1, \cdots, n$ denote standard n-dimensional unit basis vectors, points $\mathbf{x}_+ = \mathbf{x}_k \pm \Delta_k \mathbf{e}_i$ are successively examined to determine if $F_0(\mathbf{x}_+) < F_0(\mathbf{x}_k)$. Figure 1 illustrates a set of such points among which $\mathbf{x}_+$ is searched for when n=2. This set of points constitutes a pattern. If there is no point such that $F_0(\mathbf{x}_+) < F_0(\mathbf{x}_k)$, Δ_k is reduced and continued; otherwise, the step-length parameter is left unchanged or increased if a longer step is justified. Third, by setting $\Delta_{k+1} = \Delta_k$ and $\mathbf{x}_{k+1} = \mathbf{x}_+$, the iteration in the second step is repeated until $\Delta_k \leq \Delta_c$ is sufficiently small, where Δ_c is a preselected tolerance defined by the user. The final value of $\mathbf{x}_+$ represents the solution. This simple illustration suggests
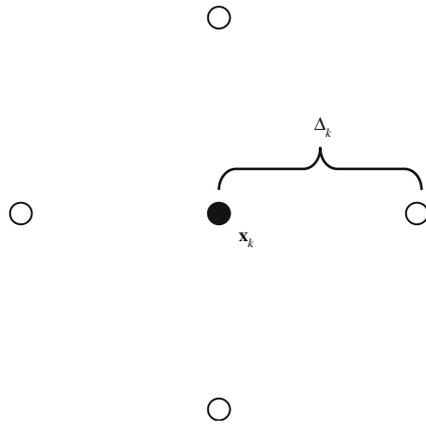
Fig. 1 A typical example of pattern search method

that there is a great deal of flexibility in pattern search algorithms, depending on how one specifies the pattern of points to be searched for in the next iteration. Despite their seeming simplicity and heuristic nature, as well as the fact that they do not have explicit recourse to the sensitivity of $F_0(\mathbf{x})$, pattern search algorithms possess global convergence properties. It has been shown that pattern search algorithms are globally convergent in unconstrained minimization, bound-constrained minimization, and linear-constrained minimization (Lewis and Torczon 2000).

If the objective F_0 is continuously differentiable, then a subsequence of the iteration produced by a pattern search method for bound and linearly constrained minimization converges to a Karush–Kuhn–Tucker point. For linear constrained problems, the geometry of nearby constraints has to be taken into account. If the problem is not at a constrained stationary point, at least one feasible descent direction with a sufficiently long distance needs to be ensured. However, it is challenging to design such types of patterns. For a general constrained problem, the augmented Lagrangian approach is developed (Lewis and Torczon 2002). The formulation assumes that any general inequality constraints have been converted into equality constraints by the introduction of non-negative slack variables. The algorithms include inner iteration, updated Lagrange multiplier estimates, and reduced penalty parameters. However, numerical study and application still appear rarely in the literature.

Although direct search methods have a special niche in modern optimization, two issues have often been cited as unavoidable concerns for any direct search method. No one would disagree that direct search is slow. Direct search methods are, in a precise sense, asymptotically slower than the steepest descent method. The other concern is the limitation on the problem size. Even in the early days of direct search methods, they were best suited for problems with a small number of variables. To overcome this drawback, this study will extend the existing pattern search algorithm by introducing the preselected response surface model to find the descent direction for iteration. This is a proper choice to set up a connection of pattern search with the response surface method as an additional trial step.

2.3 Hybrid optimization method

The goal of the optimization process is to reach a feasible minimum point for the objective function. If $\mathbf{x}_*^{(k)}$ is not a minimum point, another point $\mathbf{x}_*^{(k+1)}$ with a smaller objective function value should be obtained. In conventional RSM-based optimization procedure, reduction of the subregion size is initiated when $\mathbf{x}_*^{(k)}$ is located inside the subregion of the kth iteration, or the accuracy condition of (4) does not hold. To increase the accuracy of the response surface, a smaller subregion size is needed. In this way, a preselected convergence condition could be reached before an actual optimum point is obtained. Moreover, many original function evaluations are also added to the resulting optimization process because of the smaller subregion size of each iteration. Alexandrov et al. (1998) used a criterion based on the ratio of predicted improvement in the objective (or augmented objective for constrained problems) for managing the use of approximation models in optimization. The size of the subregion for the iteration is based on the predicted improvement. Essentially, it is a criterion based on the accuracy of the approximation. During optimization procedure especially at early stage, the size of subregion needs to be maintained big enough to locate the optimum area quickly. In other words, predicting descent, i.e., simple decrease is the focus rather than the quality of the approximations. In the proposed hybrid method, the size of each subregion is significantly reduced only when necessary, i.e., when no simple decrease is found in the iteration, which is a necessary condition for the convergence property of the pattern search method. During each iteration of the hybrid method, a numerical experiment is conducted at $\mathbf{x}_*^{(k)}$ and compared with original function values at all sample points in the feasible domain. By performing this experiment, an explicit descent condition is implemented to avoid the infeasible region and further improve the reduction of the objective function. By imposing the descent condition in the hybrid method, premature convergence and slow optimization can be avoided.

The proposed method is outlined as follows. A flow chart is shown in Fig. 2.

Step 1: Set the initial conditions and convergence criteria by defining the following parameters: (1) initial design vector $\mathbf{x}^0$, (2) sample pattern, (3) accuracy criteria $r_j^{(k)} \leq \varepsilon_j$, (4) minimum size of subregion Δ_c for convergence, and (5) convergence criteria. Set iteration number $k=0$.

Step 2: The iteration is terminated when the following convergence criteria are satisfied: (1) (4) is satisfied for all $j = 0, 1, \cdots, l$, (2) none of the move limits are active, and (3) the subregion has reached a required
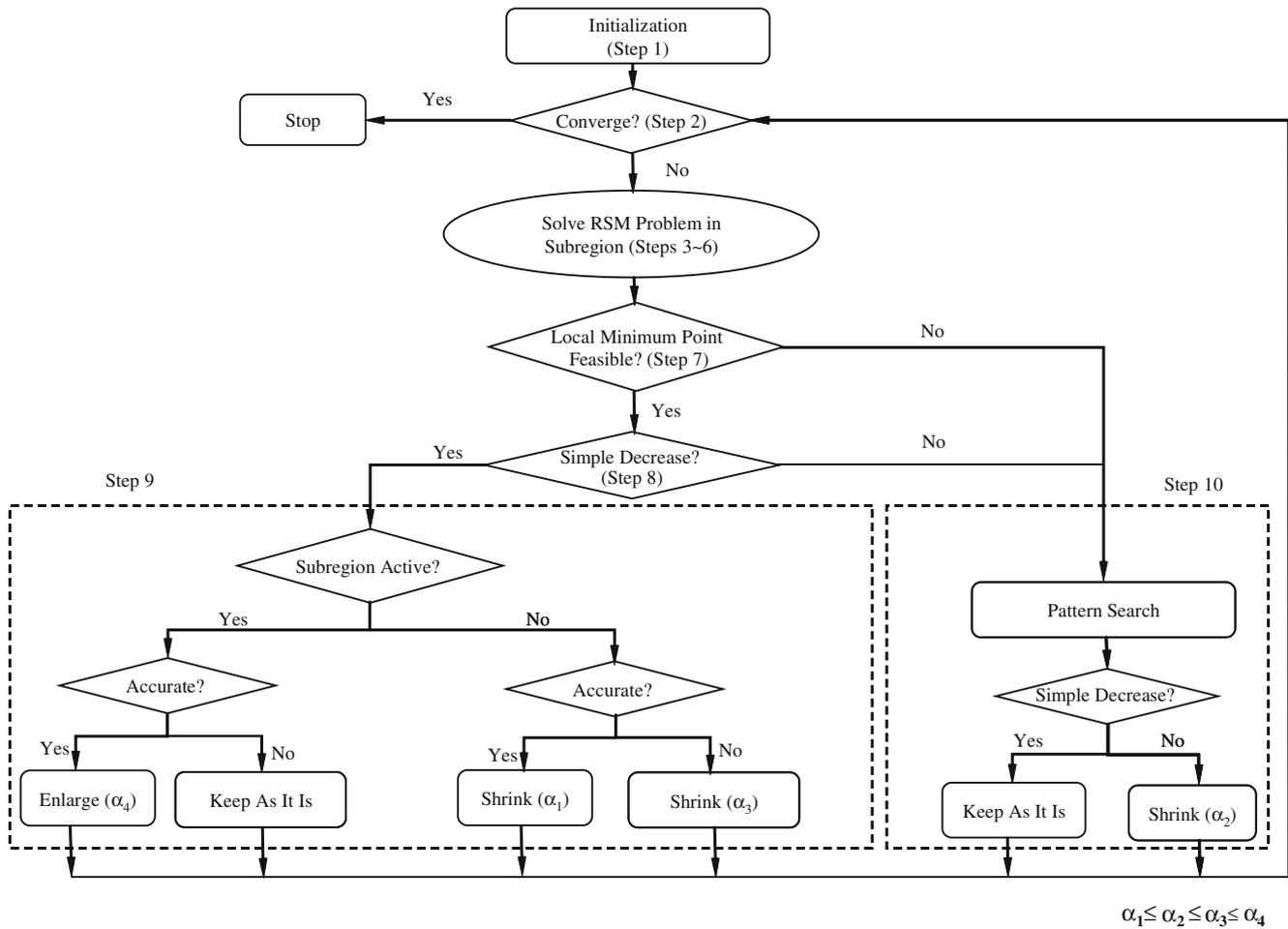
Fig. 2 Flow chart of the hybrid method

size. If convergence criteria are satisfied, then stop. Otherwise, let $k=k+1$ and go to the next step.

Step 3: Evaluate the objective and constraint function values at all sample points, as well as the gradients of objective and constraint functions if they can be readily obtained. The sampling method, which is used for function approximation and optimization, is coordinate sampling (axial star, $2n+1$).

Step 4: Determine the coefficients for RBF interpolation (or RBF plus polynomial terms in a gradient-enhanced function for the approximation problem) by solving a set of linear system equations. This step needs to be performed for all objective and constraint functions.

Step 5: Develop an RBF interpolation (or RBF plus polynomial-based MLS approximation) of objective and constraint functions.

Step 6: Solve the associated approximated problem by conventional gradient-based optimization method.

Step 7: Conduct numerical experiment at the optimum point obtained in step 6 to check feasibility and record the objective function and all constraint function values. If it is feasible, go to step 8. Otherwise, go to step 10 (pattern search) to find a new trial point (minimum feasible point from pattern points) using pattern search.

Step 8: Compare the actual objective function value at $\mathbf{x}_*^{(k)}$ with those obtained at all feasible sample points in a given pattern. The minimum feasible point becomes the new trial point. If the new trial point with simple decrease is predicted by the response surface function, then go to step 9 to check if the subregion is active. Otherwise, go to step 10 to use pattern search to locate minimum feasible point as the new trial point.

Step 9: If the subregion is active (i.e., the new trial point is located on the boundary of the subregion), then check the accuracy of the response surface approximations. If the accuracy of the response surface approximation is satisfied, then the size of the subregion for the next iteration is enlarged, otherwise, keep the size of the subregion as it is. On the
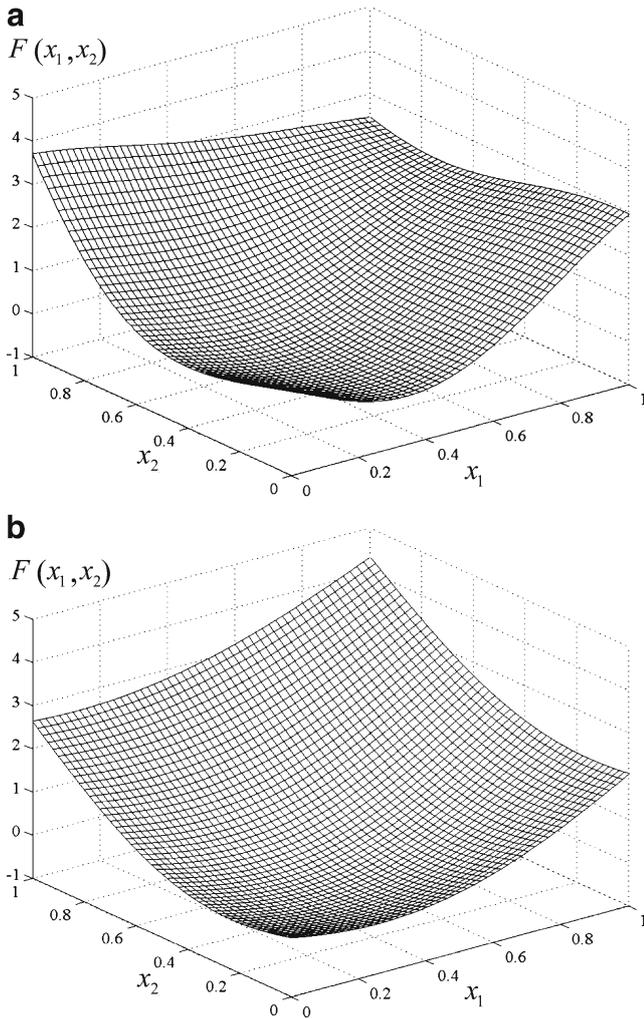
a

$F(x_1, x_2)$

b

$F(x_1, x_2)$

Fig. 3 Plots of the peak function (**a**) and exact function (**b**) (zeroth-order) by RBF

a

$F(x_1, x_2)$
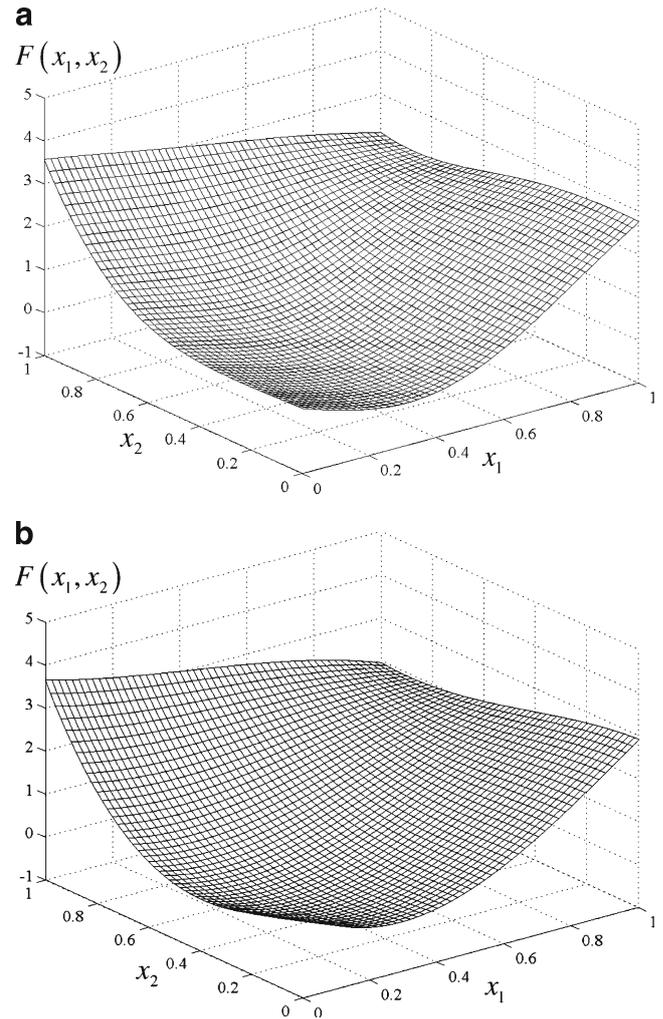

b

$F(x_1, x_2)$

Fig. 4 Plots of approximation by **a** RBF+polynomial-based MLS and **b** polynomial-based MLS (first-order)

other hand, if the subregion is not active (i.e., the new trial point is located inside the subregion), then the size of the subregion for the next iteration should be shrunk. Under this circumstance, if the response surface approximation is accurate, then the size of the subregion for the next iteration can be shrunk aggressively, compared to the case if the response surface approximations are not accurate. Set up a new pattern and go to step 2.

Step 10: The sampling points, which are the axial star, are used for the pattern search. If the new trial point obtained by the pattern search is a simple decrease, then the size of the subregion should be kept as it is. On the other hand, if a new trial point with a smaller objective function value cannot be found in this iteration, shrink the size of the new subregion according to the pattern search. Set up a new pattern and go to step 2.

3 Numerical examples

3.1 Example 1—function approximation

Consider a two-dimensional peak function (MATLAB 2004) defined as

$$F(x_1, x_2) = 3(1 - x_1)^2$$

$$\exp\left[-x_1^2 - (x_2 + 1)^2\right] - 10\left(\frac{1}{5}x_1 - x_1^3 - x_2^5\right)$$

$$\exp\left[-x_1^2 - x_2^2\right] - \frac{1}{3}$$

$$\exp\left[-(x_1 + 1)^2 - x_2^2\right],$$

$$0 < x_1 < 1, \ 0 < x_2 < 1 \tag{20}$$

Table 1 Error of response with five samples, percent [%]

	RBF	MLS	
		RBF+poly[a]	Poly[b]
Function, f	29.21	6.85	8.8
$\partial f/\partial x_1$	55.09	15.71	27.73
$\partial f/\partial x_2$	47.87	16.78	13.59

[a]*RBF+poly* RBF+polynomial

[b]*Poly* Polynomial

The zeroth-order interpolation and first-order enhanced function approximations are studied. Because response function accuracy is an important aspect for either global or local sequential approximation optimization, various function approximations are studied and compared using the same error measurement ε, defined by

$$\varepsilon = \sqrt{\frac{\int_\Omega \left[F(x_1, x_2) - \tilde{F}(x_1, x_2) \right]^2 d\Omega}{\int_\Omega F(x_1, x_2)^2 d\Omega}} \times 100 \qquad (21)$$

where $F(x_1, x_2)$ is the exact value of the function, and $\tilde{F}(x_1, x_2)$ is the response surface approximation of the function. For the zeroth-order interpolation, RBF is implemented. Because the coordinate sampling (axial star, $2n+1$) is used, for two-dimensional problem, there are five sample points. For the first-order enhanced approximation, MLS using conventional polynomial bases (cubic) and proposed RBF (c is equal to the minimum distance of the samples multiplied by factor 1.1, the same value is used in optimization) plus polynomial bases are employed to verify the accuracy. Figure 3a shows

the exact function plot in the given space. The zeroth-order interpolation plot is displayed in Fig. 3b. Although zeroth-order RBF can capture the overall character of the peak function, the amount of error at the four corners is still large. The three-dimensional plots of approximations with sensitivities, using proposed RBF plus polynomial MLS and conventional polynomial-based MLS, are given in Fig. 4a,b.

Table 1 lists numerical errors given for RBF, RBF plus polynomial-based MLS, and polynomial-based MLS. A considerable improvement is gaine*d over the interpolation based exclusively on function values. The proposed RBF plus polynomial-based MLS shows better accuracy for this example.

The example presented shows that a considerable improvement is gained by including sensitivities and using the proposed procedure over zeroth-order interpolation. Both RBF plus polynomial-based MLS and polynomial-based MLS performed well in approximation when incorporating sensitivities. It can be seen that both RBF plus polynomial-based MLS and polynomial-based MLS lead to similar accuracy levels in response surface approximation. The advantage of RBF plus polynomial-based MLS is that it provides a more systematic way of choosing bases, while conventional polynomial-based MLS requires user's experience and understanding of the nature of the system, which can be somewhat difficult in some applications.

3.2 Example 2—constrained problem

A constrained problem in two-dimensional space taken from Schittkowski (Schittkowski 1987) was originally formulated
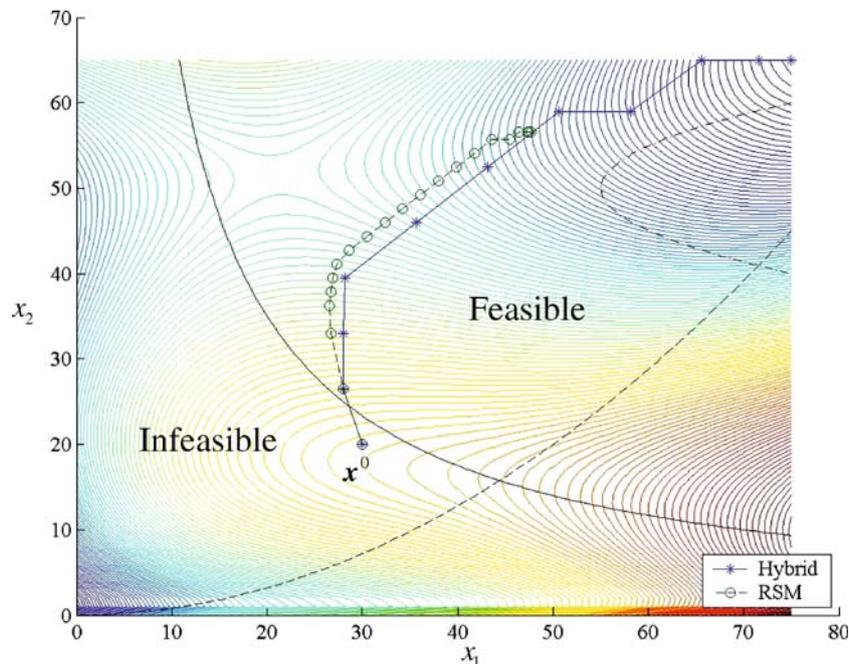

Fig. 5 Iteration history by conventional response surface and hybrid method

Table 2 The history of objective function for the constrained problem

Iteration	Conventional response surface method (RBF)	Hybrid method[a] (RBF)	
	Objective history	Objective history	Active information
0	24.479	24.479	
1	18.851	18.851	RSM
2	10.278	10.261	PS
3	5.785	1.334	RSM
4	3.610	−8.816	RSM
5	1.519	−19.123	RSM
6	−0.489	−28.633	PS
7	−2.562	−38.620	RSM
8	−4.799	−48.898	PS
9	−7.074	−56.136	RSM
10	−9.360	−58.928	PS
11	−11.633	−58.928	PS
$\vdots$			
18	−24.640		
19	−24.774		
20	−24.804		
Total analyses	101	56	

[a]*RSM* Response surface method, *PS* pattern search

by G. Barnes. It is highly nonlinear and therefore suitable as a test problem. The optimization problem is defined to

minimize $F_0(\boldsymbol{x}) = -75.196 + 3.8112x_1 - 0.12694x_1^2$

$$+ 2.0567 \times 10^{-3}x_1^3 - 1.0345 \times 10^{-5}x_1^4$$
$$+ 6.8306x_2 - 0.030234x_1x_2$$
$$+ 1.28134 \times 10^{-3}x_2x_1^2$$
$$- 3.5256 \times 10^{-5}x_2x_1^3$$
$$+ 2.266 \times 10^{-7}x_2x_1^4 - 0.25645x_2^2$$
$$+ 3.4604 \times 10^{-3}x_2^3 - 1.3514 \times 10^{-5}x_2^4$$
$$+ \frac{28.106}{x_2 + 1} + 5.2375 \times 10^{-6}x_1^2x_2^2$$
$$+ 6.3 \times 10^{-8}x_1^3x_2^2 - 7 \times 10^{-10}x_1^3x_2^3$$
$$- 3.4054 \times 10^{-4}x_1x_2^2$$
$$+ 1.6638 \times 10^{-6}x_1x_2^3$$
$$+ 2.8673 \exp(0.0005x_1x_2)$$

subject to $F_1(\mathbf{x}) = -x_1x_2/700 + 1 \leq 0$

$$F_2(\mathbf{x}) = -x_2 + 5\left(\frac{x_1}{25}\right)^2 \leq 0$$
$$F_3(\mathbf{x}) = -(x_2 - 50)^2/100 + (x_1 - 55)/20 \leq 0$$

and

$$0 \leq x_1 \leq 75$$
$$0 \leq x_2 \leq 65 \tag{22}$$

Due to the limitation in developing patterns for nonlinear problems, the conventional response surface method and the proposed hybrid method are employed. The solution to this problem is $\boldsymbol{x}_* = [75, 65]^T$, with the optimum cost value of $F_0(\boldsymbol{x}_*) = -58.928$. For both methods, optimization starts from an infeasible point $\boldsymbol{x}^0 = [30, 20]^T$, as illustrated in Fig. 5 and Table 2. In the conventional response surface method (history marked as circle), the optimum result is $F_0(\boldsymbol{x}) = -24.804$ which is premature and inaccurate. On the other hand, in the proposed hybrid method (history marked with star), optimization iteration takes advantage of the prediction of the RBF, as well as the information of sample points around the staring point to reduce the objective value. From Table 2, it is seen that RSM and pattern search method are activated during the optimization in the hybrid

Table 3 The history of objective function for the constrained problem (first-order)

Iteration	Function value (zeroth-order)[a]		Function value and sensitivities (first-order)	
	Objective (RBF)	Active[a] information	Objective[b] (RBF+poly)	Active[a] information
0	24.479		24.479	
1	18.851	RSM	18.733	RSM
2	10.261	PS	9.685	RSM
3	1.334	RSM	−5.252	PS
4	−8.816	RSM	−11.783	PS
5	−19.123	RSM	−32.450	RSM
6	−28.633	PS	−51.292	RSM
7	−38.620	RSM	−58.928	RSM
8	−48.898	PS	−58.928	PS
9	−56.136	RSM		
10	−58.928	PS		
11	−58.928	PS		
Total analyses	56		49	

[a]*RSM* Response surface method, *PS* pattern search
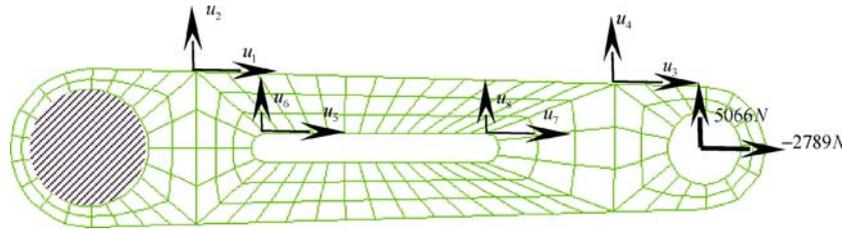[b]*RBF+poly* RBF+polynomial

Fig. 6 Design parameters of torque arm

method and yield fast (56 function analyses) and accurate solution.

In some cases, it is desirable to incorporate sensitivities to construct the response surface function if they can be obtained readily. Hence, the gradient-enhanced function approximation in optimization is also studied (the computational effort for generating the gradient is assumed to be 10% of that for function evaluation.). To have a consistent comparison, the same test problem is employed and the same coordinate sampling is also utilized. The history of objective functions is presented in Table 3 when sensitivity information is available. It is also observed that RSM and pattern search method are activated during the optimization iterations. From Table 3, it is noted that the objective value reduces rapidly when the first-order enhanced function approximations are used. The optimization procedure using gradient-enhanced approximation shows a rapid convergence rate, as compared to zeroth-order response function-based optimization.

3.3 Example 3—structural optimization

A torque-arm design optimization problem is solved by the proposed hybrid method without sensitivities. Eight shape design parameters (u_i, $i=1,..,8$) are chosen to perturb the outer/inner boundary curves of the torque arm. The shape and design parameters are shown in Fig. 6. The domain of the torque arm is discretized by 657 nodes and 177 plane stress elements. The boundary condition is imposed to fix the left hole. Young's modulus E, Poisson's ratio ν, the yield stress S_y, density ρ, and thickness t are given in Table 4. The design optimization problem is formulated such that the total mass of the structure is to be minimized, with design constraints defined for the second invariant of stress tensors (von Mises stress). Mathematically, the problem is to

minimize $\qquad\qquad m(\mathbf{u})$

subject to $\qquad\qquad \sigma_e \leq 800\,\text{MPa}$

$$4.303\,\text{mm} \leq u_1 \leq 8.303\,\text{mm}$$
$$4.709\,\text{mm} \leq u_2 \leq 6.209\,\text{mm}$$
$$34.84\,\text{mm} \leq u_3 \leq 36.84\,\text{mm}$$
$$1.68\,\text{mm} \leq u_4 \leq 5.38\,\text{mm}$$
$$6.5\,\text{mm} \leq u_5 \leq 13.0\,\text{mm}$$
$$0.0\,\text{mm} \leq u_6 \leq 3.0\,\text{mm}$$
$$26.0\,\text{mm} \leq u_7 \leq 34.0\,\text{mm}$$
$$0.0\,\text{mm} \leq u_8 \leq 2.0\,\text{mm} \qquad (23)$$

The total mass $m(\boldsymbol{u})$ of the structure is computed from the area of the structure multiplied by the thickness and density. Shape design parameters at the initial design are $u_0 =$ [7.303, 5.209, 35.84, 4.38, 12.0, 1.0, 27.0, 1.0]. The mesh at the initial design and the stress contour are plotted in Fig. 7. The highest stress of 430 MPa (node 98) occurs at the upper left of the frame. At iteration 2, mesh distortion occurred. Remeshing was carried out to continue the optimization procedure. After remeshing, there are 713 nodes and 201 plane stress elements in the domain of the torque-arm. The history of the total mass of the structure is presented in Table 5. The highest stress at the optimum design is increased to 799 MPa (node 250 in Fig. 8), which occurs at the upper right frame. Shape design parameters at the optimum are $u_* =$ [6.896, 4.709, 35.651, 1.68, 10.211, 3.0, 28.654, 1.298]. The structural mass is reduced from 0.875 to 0.498 kg (43.02%) through optimization by the proposed hybrid method.

4 Application to microelectronic packaging system

4.1 Fatigue durability analysis of microelectronic packaging systems

The low-cycle fatigue is a common failure mechanism in solder joints in the electronic packaging systems. The cyclic
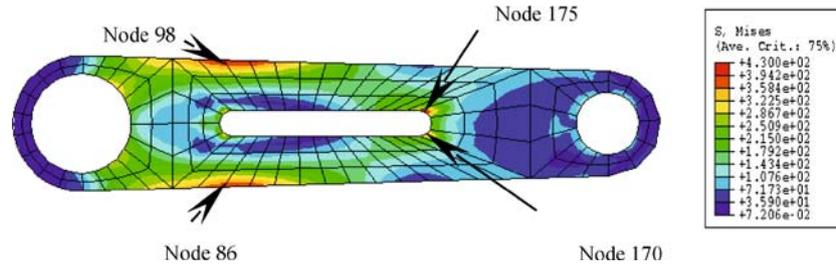
Table 4 Material properties of torque arm

Property	Value
E	207 GPa
ν	0.3
S_y	800 MPa
ρ	7,800 kg/m^3
t	3 mm

Fig. 7 Initial stress distribution

thermal loading combined with the mismatch in thermal expansion properties of various components of the assembly leads to stress reversals and accumulation of inelastic strain in solder joints. Because the solder material exhibits creep deformation characteristics, nonlinear viscoelastic or viscoplastic finite element analysis (FEA) is often needed to calculate the fatigue life. However, such a detailed analysis can become computationally demanding and even prohibitive. To overcome this problem, The University of Iowa and Rockwell Collins team developed a novel global–local methodology to predict mechanical deformation and fatigue durability of solder joints for a given design (Zhang et al. 2006). The methodology involves three major steps: (1) a global deformation analysis employing an optimal geometry of equivalent solder joints derived from nonlinear load-deformation response, (2) a local critical solder-joint analysis involving rigorous application of nonlinear submodeling technique, and (3) a fatigue life analysis including fatigue crack-initiation and/or crack-propagation. Further details of the methodology and its experimental validation are available in the work of Zhang et al. (2006).

4.1.1 ASAT 144 fpBGA packaging system and initial design

The global–local analysis methodology was applied to the ASAT 144 fine-pitch ball-grid array (fpBGA) package, as

Table 5 Objective history of the total mass

Iteration	Total mass	Active information[a]
0	0.875	
1	0.746	RSM
2	0.646	RSM
3	0.590	PS
4	0.552	RSM
5	0.532	PS
6	0.514	RSM
7	0.503	RSM
8	0.498	RSM
9	0.498	PS
10	0.498	PS
Total analyses	171	

[a]*RSM* Response surface method, *PS* pattern search

shown in Fig. 9. The printed circuit board is made of FR4 material with its time-dependent elastic material properties defined in Table 6. The time-independent elastic material properties of package components (rigid carrier, die chip, die attachment, and molding compound) and solder are defined in Table 7. The solder material is eutectic 63Sn-37Pb, which follows a hyperbolic-sine law (Pan 1991). The thermal load profile for both packages involves temperature cycles varying from −55 to 125°C.

Figure 10a displays a global model (quarter) of the fpBGA package. The FEA discretization by ABAQUS (version 6.5) (ABAQUS 2005) consists of 4,663 eight-noded solid elements and 6,408 nodes. All solder joints were replaced by an equivalent two-parameter diamond model (Zhang et al. 2006). Figure 10b shows a local model of the critical solder joint that consists of 2,432 eight-noded solid elements and 2,707 nodes.

For the initial design, the fatigue crack-initiation life, calculated using the local model (Fig. 10b) and Coffin–Manson equation (Zhang et al. 2006), was predicted to be 144 cycles. Figure 11 presents the fatigue-life contour plots of the solder joints for crack initiation. The top view is given in Fig. 11a, and the bottom view is given in Fig. 11b. Figure 12 shows the plot of global relative displacements $\Delta u_i, i = 1,2,3$, which represents the global relative displacements in the x, y, and z directions for three thermal cycles.

4.2 Optimization of ASAT 144 fpBGA packaging system

It is well known that thermal cyclic failure is a serious concern in the electronic packaging system. Under temperature cycles, a mismatch of the coefficients of thermal expansion (CTE) in the assembly can induce repeated stresses, resulting in fatigue damage accumulation in the solder joints. The objective is to select proper mechanical properties of the package and solder-joint geometrical parameters to improve solder-joint fatigue life and thus obtain a more reliable design. To resolve this issue, optimization of the whole assembly of microelectronic packaging becomes necessary. In general, the design parameters of such complicated electronic packages are CTE of assembly materials, and solder ball shape parameters. The optimization problem can be formulated to maximize fatigue crack-initiation life with respect to the de-
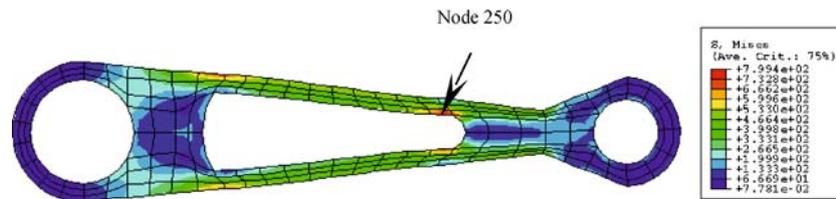
Fig. 8 Stress distribution at optimum design

sign variables i.e., the selected parameters. Mathematically, it is to

$$\text{maximize } N_f(\|\Delta \boldsymbol{u}(\boldsymbol{\alpha}; \boldsymbol{D})\|, \boldsymbol{D})$$

$$\text{subject to: } \begin{array}{l} \boldsymbol{\alpha}_l \leq \boldsymbol{\alpha} \leq \boldsymbol{\alpha}_u \\ \boldsymbol{D}_l \leq \boldsymbol{D} \leq \boldsymbol{D}_u \end{array} \qquad (24)$$

where $\boldsymbol{\alpha}$ is the vector of selected material properties (CTE), $\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_u$ are the lower and upper bounds, respectively, $\boldsymbol{D}$ is the vector of solder ball shape parameters, and $\boldsymbol{D}_l$ and $\boldsymbol{D}_u$ are the lower and upper bounds, respectively. The exact optimum solution of the packaging system should be obtained directly from (24). However, due to the com-
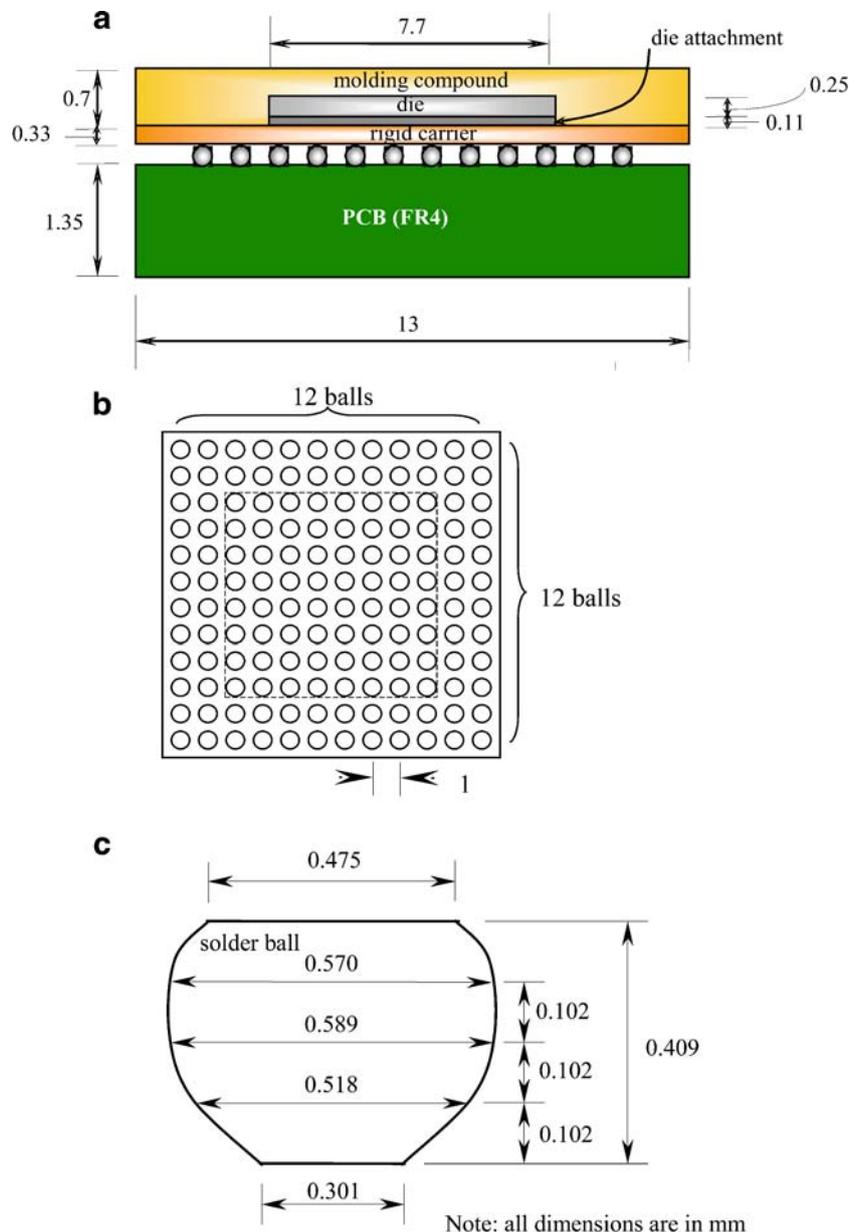

Fig. 9 ASAT 144 fpGBA package **a** front view, **b** top view, **c** solder ball shape

Table 6 Elastic properties of FR4 material[a]

Properties	Temperature (°C)			
	-55	30	95	125
E_x, GPa	22.4	22.4	20.7	19.3
E_y, GPa	22.4	22.4	20.7	19.3
E_z, GPa	1.6	1.6	1.2	1.0
ν_{xy}	0.02	0.02	0.02	0.02
ν_{yz}	0.143	0.143	0.143	0.143
ν_{yz}	0.143	0.143	0.143	0.143
G_{xy}, GPa	0.63	0.63	0.6	0.5
G_{yz}, GPa	0.199	0.199	0.189	0.167
G_{zx}, GPa	0.199	0.199	0.189	0.167
α_x, /°C	15.85×10^{-6}	15.85×10^{-6}	15.85×10^{-6}	15.85×10^{-6}
α_y, /°C	19.14×10^{-6}	19.14×10^{-6}	19.14×10^{-6}	19.14×10^{-6}
α_z, /°C	80.46×10^{-6}	80.46×10^{-6}	80.46×10^{-6}	80.46×10^{-6}

(a) Symbols E, ν, α, and G denote elastic modulus, Poisson's ratio, coefficients of thermal expansion, and shear modulus, respectively. The subscripts indicate components for orthotropic properties.

plicated nature of the problem, it is generally prohibitively expensive to perform optimization in a single procedure. A more effective alternative is proposed, which implements a two-stage optimization approach: (1) optimization of global relative displacements and (2) shape optimization of local solder joint. This decomposition assumes that the changes of local solder ball shape do not significantly affect the global relative displacements, which is dominantly controlled by the thermal mismatch of the packaging assembly. In the conventional, gradient-based optimization procedure, sensitivities (derivatives) with respect to design parameters should be obtained. Because of the nonlinearity, creep behavior, and path-dependent nature of the electronic packaging analysis, it is very difficult to calculate design sensitivities. In this paper, the proposed hybrid method is applied to the two-stage optimization process.

4.2.1 Optimization of global relative displacements

Essentially, shorter fatigue life of the solder ball is due to large relative deformations of the packaging systems. Because the package and PCB transmit the displacement as the boundary condition to the local model, the global relative displacements

Table 7 Other material properties of package and solder ball

Material	Young's modulus (E), GPa	Poisson's ratio (ν)	Coefficient of thermal expansion (α), /C
Solder	24.83	0.4	21×10^{-6}
Rigid carrier	26	0.39	13×10^{-6}
Die attachment	16	0.25	14×10^{-6}
Die chip	131	0.3	4.1×10^{-6}
Molding compound	16	0.25	14×10^{-6}
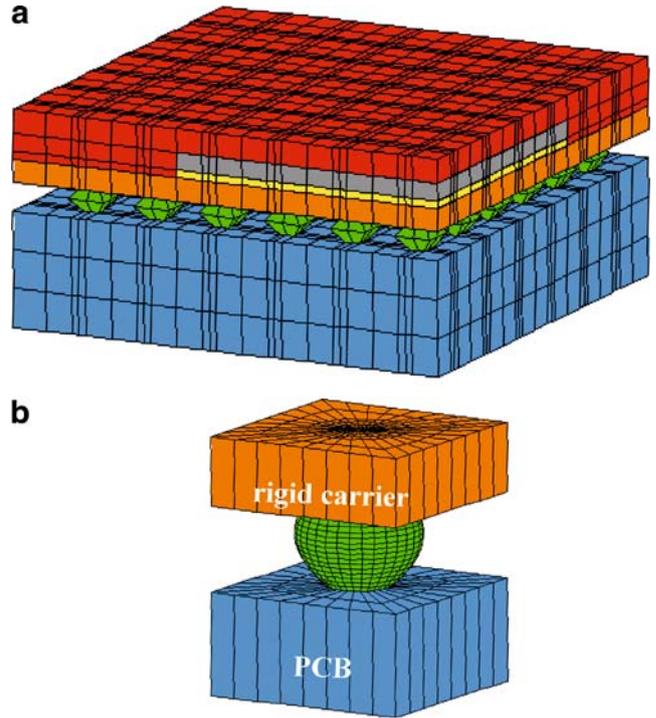
Fig. 10 A schematic of the global–local methodology **a** global model and **b** local model

are selected in the first stage as the objective function, with the CTE of materials selected as design variables. Because the components of the flip chip assembly have different CTE, the thermomechanical fatigue failure of the solder joints will occur under cyclic thermal loading. CTE of various components such as PCB, rigid carrier, and molding compound are considered as design variables in this study. Different combinations of design variables are studied for the ASAT 144 fpBGA packaging system. The optimization formulation is to

$$\text{minimize } \|\Delta \mathbf{u}(\boldsymbol{\alpha}; \mathbf{D}_0)\|^2$$
$$\text{subject to: } \boldsymbol{\alpha}_l \leq \boldsymbol{\alpha} \leq \boldsymbol{\alpha}_u \tag{25}$$

where $\Delta u_i = u_{i,top} - u_{i,bottom}$ ($i=1,..,3$) represents the relative displacements in the x, y, and z directions; $\boldsymbol{\alpha}$ are the CTE of materials, which are considered as design variables; $\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_u$ are the lower and upper bounds of material properties, respectively; and $\boldsymbol{D}_0$ is the parameter of initial solder ball shape.

Different cases of combination of design variables, which are defined in Table 8, are studied. In case G1, the CTE of PCB (in x, y, and z directions) are selected as design variables; in case G2, the CTE of rigid carrier and molding compound in the package are selected as design variables; for case G3, the CTE of PCB (in x, y, and z directions), rigid carrier, and molding compound of the whole electronic packaging assembly are considered.

Case G1 In this case, CTE of PCB are selected as design variables. PCB is modeled as an orthotropic material

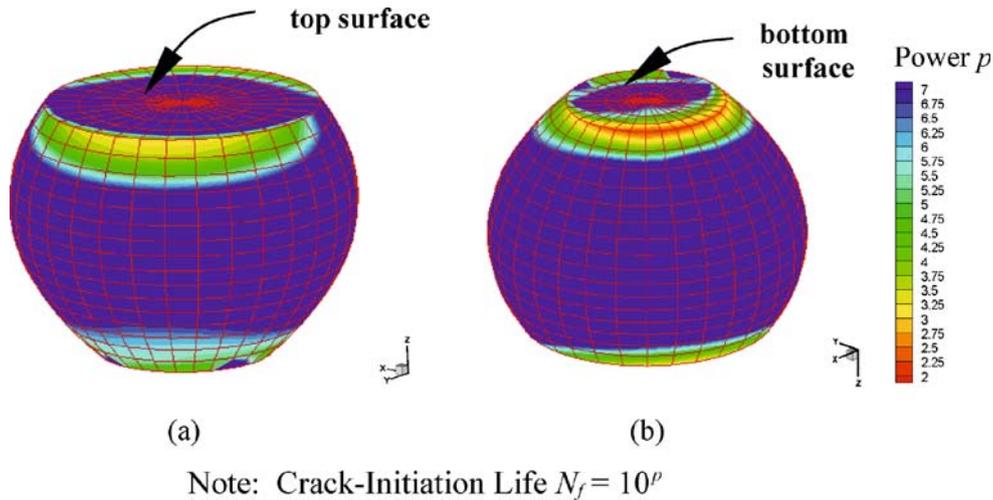Note: Crack-Initiation Life $N_f = 10^p$

Fig. 11 The strain-based crack-initiation life contour plots at the initial design **a** top view and **b** bottom view

with mechanical properties that exhibit dependence on temperature. The design bounds of α_x, α_y, and α_z in three directions are defined as $11.10 \times 10^{-6} \leq \alpha_x \leq 20.62 \times 10^{-6}$, $13.40 \times 10^{-6} \leq \alpha_y \leq 24.88 \times 10^{-6}$, and $56.32 \times 10^{-6} \leq \alpha_z \leq 104.6 \times 10^{-6}$.

Applying the proposed hybrid optimization method, it is found that the global relative displacements are reduced when CTE in the x and y directions are reduced. The coefficient of thermal expansion in the z direction has no significant influence to the global relative displacement. Figures 12 and 13 show the global relative displacement histories at the initial design and optimum design of CTE of PCB, respectively. At the optimum design, the CTE values are [12.0178×10^{-6}, 13.5302×10^{-6}, 80.4496×10^{-6}], and at the initial design, they are [15.85×10^{-6}, 19.14×10^{-6}, 80.46×10^{-6}]. It can be seen that the global relative displacements in x and y directions are reduced significantly, while the global relative
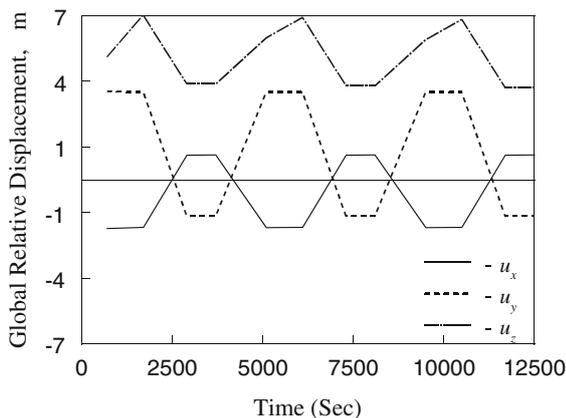
displacement in z direction exhibits almost no change. The fatigue life for the initial CTE of PCB is 144 cycles, while the fatigue life for the optimum CTE of PCB is 2,756 cycles, as shown in Table 8, a 19-time improvement. The optimum results also show that the global relative displacements in x and y directions play a critical role in affecting the fatigue life.

Case G2 In this case, CTE of package including both of rigid carrier (α_{rc}) and molding compound (α_{mc}) are selected as design variables. They are modeled as isotropic materials. The design bounds of α_{rc} and α_{mc} are defined as $9.10 \times 10^{-6} \leq \alpha_{rc} \leq 16.9 \times 10^{-6}$ and $9.00 \times 10^{-6} \leq \alpha_{mc} \leq 20.0 \times 10^{-6}$.

At the optimum design, the CTE values are [14.8576×10^{-6}, 20.0000×10^{-6}], and at the initial design are [13.0×10^{-6}, 14.0×10^{-6}]. It is noted that α_{rc} and α_{mc} need to be increased to reduce global relative displacements. Figure 14 shows the global relative displacement histories at the optimum of CTE of the package. Compared to Fig. 12, it can be observed that the global relative deformation in the z direction is reduced significantly, and the global relative displacements in the x and y directions have also decreased. The fatigue life at the initial CTE of the package is 144 cycles, whereas the fatigue life at the optimum CTE of the package is 1,914 cycles, as shown in Table 8, a 13-time improvement. The optimum results show that the global relative deformation in z direction does not play a significant role in determining the fatigue life as the x and y direction displacements.

Case G3 In this case, CTE of the assembly (including PCB, rigid carrier, and molding compound) are selected as design variables. The same design bounds as in case G1 and case G2 are applied here.

Utilizing the hybrid optimization method, the optimum CTE values are obtained as [11.896×10^{-6}, 13.4×10^{-6}, 80.4532×10^{-6}, 13.5971×10^{-6}, 18.7216×10^{-6}]. Figure 15 shows the global relative displacement histories at the op-

Fig. 12 The global relative deformation at initial design

Table 8 Case study for optimization of global displacements

Optimization of Global Model	Case: G1 (PCB)	Case: G2 (Package)	Case: G3 (PCB + Package)
Design Variables	CTE of PCB in three directions (orthotropic)	CTE of rigid carrier and molding compound (isotropic)	CTE of PCB in three directions and CTE of rigid carrier and molding compound
Initial Fatigue Life (Cycles)		144	
Design Changes (/°C)	CTE in x: 1.58531E-05 ↓ 1.20178E-05 CTE in y: 1.91400E-05 ↓ 1.35302E-05 CTE in z: 8.04566E-05 ↓ 8.04496E-05	CTE of rigid carrier: 1.300E-05 ↓ 1.48576E-05 CTE of molding compound: 1.400E-05 ↓ 2.000E-05	CTE in x: 1.58531E-05 ↓ 1.18960E-05 CTE in y: 1.91400E-05 ↓ 1.34000E-05 CTE in z: 8.04566E-05 ↓ 8.04532E-05 --- CTE of rigid carrier: 1.300E-05 ↓ 1.35971E-05 CTE of molding compound: 1.400E-05 ↓ 1.87216E-05
Optimum Fatigue Life (Cycles)	2756	1914	20483

timum CTE of the assembly. Because case G3 includes all CTEs, the global relative displacements in all three directions are expected to be significantly reduced. The fatigue life at the optimum CTE of the assembly is 20,483 cycles, as shown in Table 8, a 142-time improvement, compared to the initial CTE with 144 cycles. The optimum results show that the CTE of the assembly have significant influence to the thermal fatigue life.

The summary of these case studies is given in Table 8. From the above case studies, it is noted that the CTE of materials in package have significant influence in the deformation

of z direction, while the CTE of PCB has an effect to the deformation in x and y directions.

4.2.2 Shape optimization of solder ball

In the local analysis, shape optimization of the solder ball is carried out using the global displacement response as boundary condition. For this, two extreme cases, i.e., case G1 (CTE of PCB as design variables) and case G3 (CTE of PCB, rigid carrier, and molding compound as design variables) of the
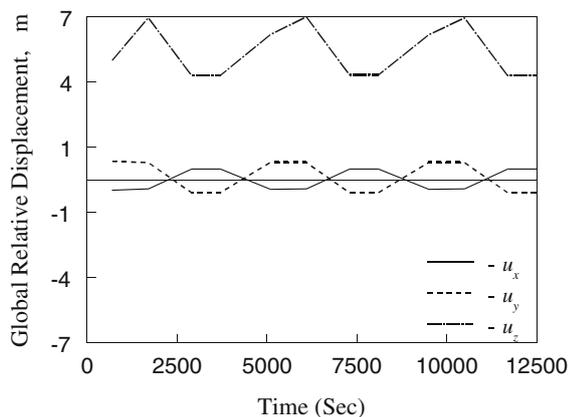

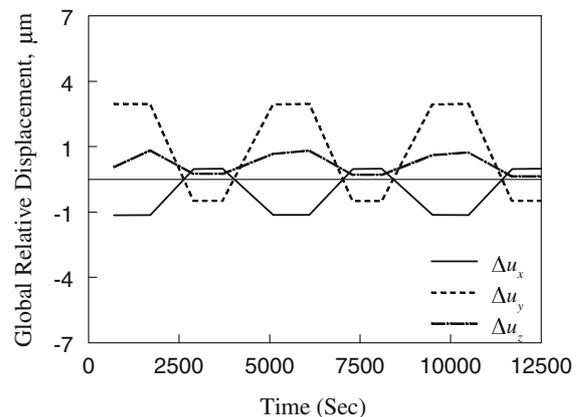**Fig. 13** The global relative deformation at optimum design of CTE of PCB

Fig. 14 The global relative deformation at optimum design of CTE of package
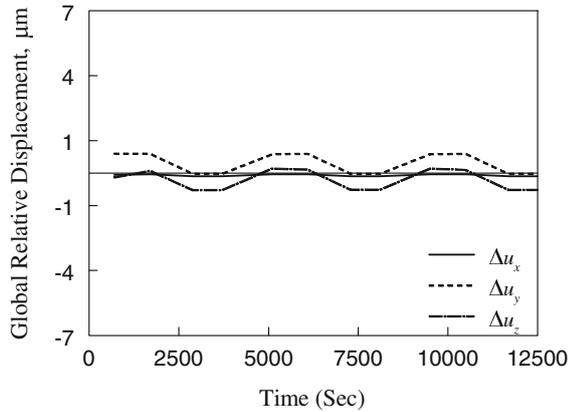
Fig. 15 The global relative deformation at optimum design of CTE of assembly

optimization of the global relative displacements, are used as input for the shape optimization of the solder ball. The solder ball shape parameters, such as the diameters of the top surface (D_1, mm), middle surface (D_2, mm), and bottom surface (D_3, mm), are selected as design variables, as shown in Fig. 16. The optimization formulation is to

$$\text{maximize } N_f(\|\Delta u(\alpha_*; D_0)\|, D)$$
$$\text{subject to: } 0.2850 \leq D_1 \leq 0.6651$$
$$0.3537 \leq D_2 \leq 0.8252$$
$$0.1803 \leq D_3 \leq 0.4207$$
$$D_1 \leq D_2$$
$$D_3 \leq D_2 \qquad (26)$$

where N_f is the fatigue crack-initiation life, α_* is the optimum results from the global model, D_0 is the initial value of design variables, D is the vector of design variables, and D_l and D_u are the vectors of lower and upper bounds, respectively, of design variables. D_2 needs to be larger than D_1 and D_3 to avoid a concave solder ball shape.

Case G1-L In this case, the solder ball shape optimization is carried out using the boundary conditions obtained from the optimum CTE of PCB in case G1. Using the proposed hybrid optimization method, the design variables D_1, D_2, and D_3 at optimum are obtained as [0.3040, 0.5187, 0.4087] mm from the initial design [0.4750, 0.5894, 0.3005] mm. Figure 17 shows the fatigue life contour at the optimum design. The diameter of the top surface decreases significantly, while the diameter of the bottom surface is increased. The fatigue life improves from 2,756 cycles to 40,736 cycles, a 15-time improvement, as shown in Table 9.

Case G3-L In this case, the solder ball shape optimization is carried out using the boundary conditions obtained from the optimum CTE of PCB, rigid carrier and molding compound in Case G3. Through design optimization, the design variables D_1, D_2 and D_3 at optimum are obtained as [0.4715, 0.5423, 0.3354] mm from the initial design [0.4750, 0.5894, 0.3005] mm. In this case, the diameters of the top, middle,

and bottom surfaces change very slightly. Figure 18 shows the fatigue life contour at the optimum design. The fatigue life is improved from 20,483 cycles to 46,899 cycles, a 2.29-time improvement, as shown in Table 9.

The summary of these case studies for solder ball shape optimization is given in Table 9. From these results, it is observed that solder ball shape parameters significantly influence durability performance of the electronic packaging system when global relative deformations due to thermal mismatch are large.

5 Conclusions

This paper presents a new hybrid optimization method and its application to find optimum designs of the microelectronic packaging system. In this method, the gradient-based algorithm and pattern search algorithm are integrated for robust and efficient optimization process. The method is based on multipoint interpolations of the objective and constraint functions, a multiquadric RBF (or RBF plus polynomial-based moving least-squares method if sensitivities are available) for function approximation, and a pattern search algorithm to impose a descent condition. Numerical examples are presented to illustrate the accuracy and computational efficiency of the proposed method for both function approximation and design optimization. A global–local optimization approach is proposed for design optimization of ASAT 144 fpBGA microelectronic packaging system to increase fatigue life. Different combinations of material properties and shape design parameters are evaluated in several case studies. It is observed that material properties of the packaging assembly and solder ball shape parameters have significant influences to the thermal fatigue life of solder joints. Through design optimization, the durability performance of the microelectronics packaging system is improved significantly.
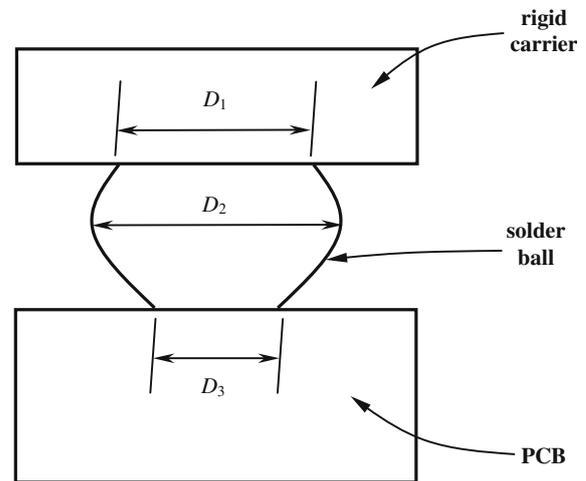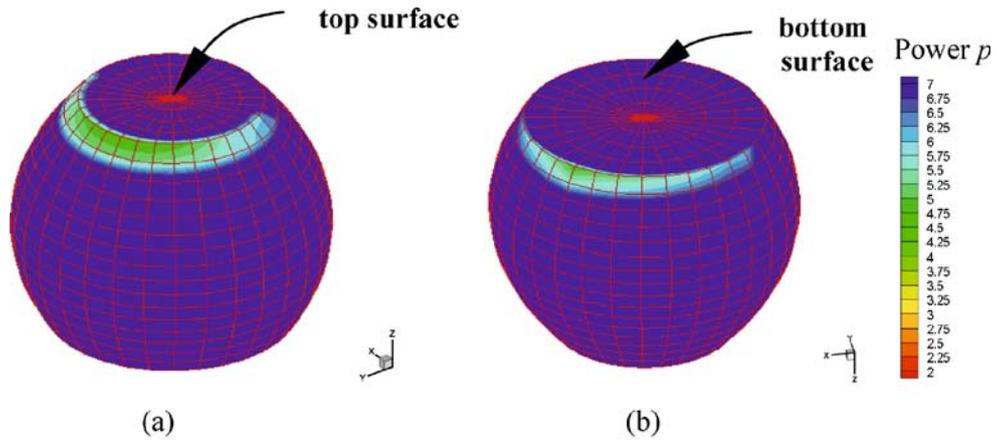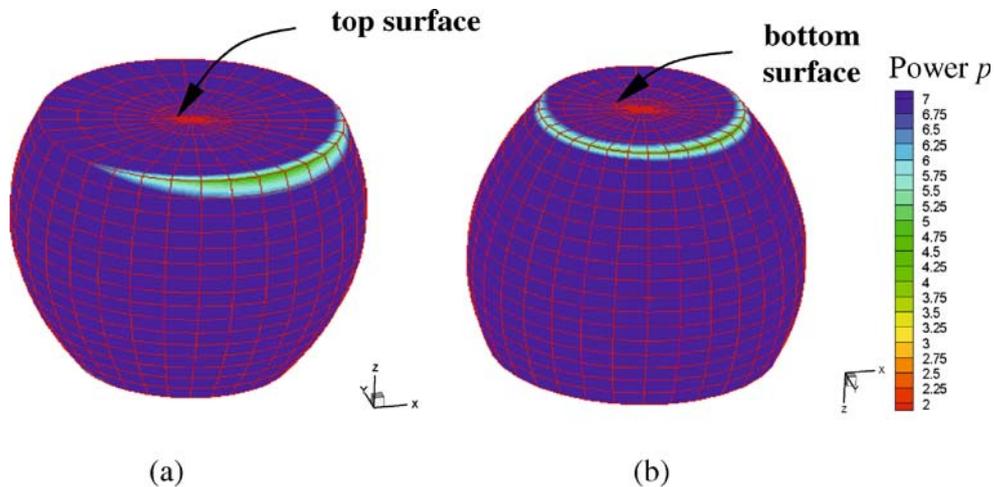

Fig. 16 Design parameters of local model

Fig. 17 The strain-based crack-initiation life contour plots at the optimum local shape design of case G1-L **a** top view and **b** bottom view

Shape optimization of solder ball	Boundary condition transformed from case G1 (PCB)	Boundary condition transformed from case G3 (PCB+package)
	Case G1-L	Case G3-L
Design variables	D_1, D_2, D_3	D_1, D_2, D_3
Initial fatigue life (global optimum, cycles)	2,756	20,483
Design changes (mm)	$D_1: 0.4750 \rightarrow 0.3040$	$D_1: 0.4750 \rightarrow 0.4715$
	$D_2: 0.5894 \rightarrow 0.5187$	$D_2: 0.5894 \rightarrow 0.5423$
	$D_3: 0.3005 \rightarrow 0.4087$	$D_3: 0.3005 \rightarrow 0.3354$
Optimum fatigue life (cycles)	40,736	46,899

Table 9 Case study for shape optimization of solder ball

Fig. 18 The strain-based crack-initiation life contour plots at the shape optimum local design of case G3-L **a** top view and **b** bottom view

References

ABAQUS (2005) User's guide and theoretical manual, version 6.5, Hibbitt. Karlsson, and Sorenson, Pawtucket, RI

Alexandrov NM, Dennis JE, Lewis,RM, Torczon V (1998) A trust-region framework for managing the use of approximation models in optimization. Struct Multidisc Optim 15:16–23

Box G (1957) Evolutionary operation: a method for increasing industrial productivity. Appl Stat 6:81–101

Choi KK, Kim NH (2002) Design optimization of springback in a deep drawing process. AIAA J 40(1):147–153

Choi KK, Kim NH (2004) Structural sensitivity analysis and optimization. Springer, Berlin Heidelberg New York

Chung H, Alonso J (2002) Using gradients to construct cokriging approximation models for high-dimensional design optimization problems. 40th AIAA aerospace sciences meeting and exhibit, Reno, NV, 14–17 January 2002

Franke R (1982) Scattered data interpolation: tests of some methods. Math Comput 38:181–200

Frank PD, Booker AJ, Caudell TP, Healy MJ (1992) A comparison of optimization and search methods for multidisciplinary design optimization. Proceedings of 4th AIAA/USAF/OAI symposium on multidisciplinary design optimization, Cleveland, OH, 21–23 September 1992

Giunta A, Dudley JM, Narducci R, Grossman B, Haftka RT, Mason WH, Watson LT (1994) Noisy aerodynamic response and smooth approximation in HSCT design. AIAA paper 94-4376-CP presented at the AIAA/NASA/USAF/ISSMO symposium. On multidisciplinary analysis and optimization, Panama City, FL

Glowinski R, Kearsley AJ (1995) On the simulation and control of some friction constrained motions. SIAM J Optim 3(5):681–694

Kim N, Choi K, Chen J (2000) Shape design sensitivity analysis and optimization of elasto-plasticity with frictional contact. AIAA J 38(9):1742–1753

Kim N, Choi K, Chen J (2001) Structural optimization of finite deformation elastoplasticity using continuum-based design sensitivity formulation. Comput Struct 79(20–21):1959–1976

Lau J (1991) Solder joint reliability—theory and applications. Van Nostrand Reinhold, NY

Lau J (1995) Ball grid array technology. McGraw-Hill

Lewis RM, Torczon VV (2000) Pattern search methods for linearly constrained minimization. SIAM J Optim 10(3):917–941

Lewis RM, Torczon VV (2002) A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. SIAM J Optim 12(4):1075–1089

Lewis RM, Torczon VV, Trosset M (1998) Why pattern search works, NASA/CR-208966, ICASE Report No. 98–57

Light W (1992) Advances in numerical analysis: wavelets, subdivision algorithms, and radial basis functions, vol 3. Clarendon, Oxford

Liu W, Batill S (2000) Noise gradient-enhanced neural network response surface approximation. AIAA paper 2000-4923, AIAA multidisciplinary analysis and optimization conference and exhibit, Long Beach, CA, September 2000

Liu B, Haftka RT, Watson LT (2004) Global–local structural optimization using response surfaces of local optimization margins. Struct Optim 27:352–359

MATLAB, version 6.5. The MathWorks

Myers R, Montgomery DC (1995) Response surface methodology: process and product optimization using designed experiments. Wiley, New York

Ohring M (1998) Reliability and failure of electronic materials and devices. Academic, New York

Pan Tsung-Yu (1991) Thermal cycling induced plastic deformation in solder joints—Part 1: accumulated deformation in surface mount joints. Transactions of the ASME J Electron Packag 113:8–15 (March)

Schittkowski K (1987) More test examples for nonlinear programming codes. Springer, Berlin Heidelberg New York

Toropov V, Filatov AA, Polynkin AA (1993) Multiparameter structural optimization using FEM and multipoint explicit approximations. Struct Optim 6:7–14

Van Keulen F, Liu B, Haftka RT (2000) Noise and discontinuity issues in response surfaces based on functions and derivatives. AIAA paper 2000–1363, Proceedings of 41th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, Atlanta, Georgia, 3–6 April 2000

Venter G, Haftka RT, Starner JH Jr (1998) Construction of response surface approximations for design optimization. AIAA 36(12): 2242–2249

Wang BP (2004) Parameter optimization in multiquadric response surface approximations. Struct Multidisc Optim 26:219–223

Wang JG, Liu GR (2002) On shape parameters for radial PIM meshless method. Comput Methods Appl Mech Eng 191:285–296

Wendland H (1995) Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. Adv Comput Math 4:389–396

Youn BD, Choi KK (2004) A new response surface methodology for reliability-based design optimization. Comput Struct 82(2–3): 241–256

Zhang X, Song KZ, Lu MW, Liu X (2000) Meshless methods based on collocation with radial basis functions. Comput Mech 26:333–343

Zhang T, Rahman S, Choi KK, Cho K, Baker P, Shakil M, Heitkamp D (2006) A global–local approach for mechanical deformation and fatigue durability of microelectronic packaging systems (in press)