Artificial Intelligence in Medicine (2007) **40**, 115–126

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

# Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment

**Jaime S. Cardoso** [a,*], **Maria J. Cardoso** [b]

[a] *Faculdade de Engenharia and INESC Porto, Universidade do Porto, Campus da FEUP, Rua Dr. Roberto Frias, no. 378 4200-465 Porto, Portugal*
[b] *Faculdade de Medicina, Universidade do Porto, Alameda do Prof. Hernâni Monteiro, 4200-319 Porto, Portugal*

Received 1 September 2006; received in revised form 19 January 2007; accepted 8 February 2007

**KEYWORDS**
Computer-aided medical system;
Breast cancer conservative treatment;
Aesthetical evaluation;
Support vector machines

**Summary**

*Objective:* This work presents a novel approach for the automated prediction of the aesthetic result of breast cancer conservative treatment (BCCT). Cosmetic assessment plays a major role in the study of BCCT. Objective assessment methods are being preferred to overcome the drawbacks of subjective evaluation.
*Methodology:* The problem is addressed as a pattern recognition task. A dataset of images of patients was classified in four classes (*excellent*, *good*, *fair*, *poor*) by a panel of international experts, providing a gold standard classification. As possible types of objective features we considered those already identified by domain experts as relevant to the aesthetic evaluation of the surgical procedure, namely those assessing breast asymmetry, skin colour difference and scar visibility. A classifier based on support vector machines was developed from objective features extracted from the reference dataset.
*Results:* A correct classification rate of about 70% was obtained when categorizing a set of unseen images into the aforementioned four classes. This accuracy is comparable with the result of the best evaluator from the panel of experts.
*Conclusion:* The results obtained are rather encouraging and the developed tool could be very helpful in assuring objective assessment of the aesthetic outcome of BCCT.


* Corresponding author. Tel.: +351 222094000; fax: +351 222094250.
*E-mail addresses:* jaime.cardoso@inescporto.pt (J.S. Cardoso), mjcard@med.up.pt (M.J. Cardoso).
*URL:* http://www.inescporto.pt/~jsc

0933-3657/$ – see front matter 
doi:10.1016/j.artmed.2007.02.007

## 1. Introduction

Breast cancer conservative treatment has been increasingly used over the last few years, as a consequence of its more acceptable cosmetic outcome when compared with mastectomy, but with identical oncological results. Although considerable research has been put into BCCT techniques, diverse aesthetic results are common, highlighting the importance of this evaluation in institutions performing breast cancer treatment, so as to improve working practices.

The categorization of the aesthetical result relies on the complex interplay of various factors, subjectively estimated and combined by observers through visual inspection. Traditionally, aesthetic evaluation has been performed subjectively by one or more observers [1–3]. The accurate evaluation of aesthetic result is mainly dependent on the observers' experience because different and complementary variables and estimations are combined synergistically in assigning an evaluation grade. Taking into account the inherent subjectivity of any human decision, the final evaluation of aesthetic result performed by observers is questionable. In fact, this form of assessment has been shown to be poorly reproducible [4–7], which creates uncertainty when comparing results between studies. It has also been demonstrated that observers with different backgrounds evaluate cases in different ways [8].

Objective methods of evaluation have emerged as a way to overcome the poor reproducibility of subjective assessment and have until now consisted of measurements between identifiable points on patient photographs [4,7,9]. The correlation of objective measurements with subjective overall evaluation has been reported by several authors [5–7,10]. Until now though, the overall cosmetic outcome was simply the sum of the individual scores of subjective and objective individual indices [5,6,10,11].

The present study introduces a new methodology for assisting the aesthetical evaluation of BCCT, based on an optimized objective score for the quantification of the aesthetic results. The proposed method lies in the cross-section of medical expert systems, soft computing and machine learning. The method discriminates among the four categories of the best-known classification.

This introduction is concluded with a brief review of the work done in this area. In the next section we describe the data used and the procedure followed to obtain a gold standard classification. Section 3 outlines the analysis techniques employed. Section 4 presents the results of the experimental investigations. Finally, conclusions of the work are given in the last section.

### 1.1. Brief history of the cosmetic assessment measures

Harris [1] introduced a subjective overall cosmetic score, that would later become a de facto standard: *excellent* (treated breast nearly identical to untreated breast), *good* (treated breast slightly different than untreated), *fair* (treated breast clearly different from untreated but not seriously distorted) and *poor* (treated breast seriously distorted).

Until Pezner [4], the reported works — of which [12,13] are examples — have focused on the correlation between the overall cosmetic result, the effects of the surgery and the technique used, with the assessment being subjectively performed by observers. With Pezner [4] the objective assessment of the cosmetic result of the surgery was introduced with the first objective measure to evaluate asymmetry, one of the aspects of cosmesis: breast retraction assessment (BRA). In Pezner [14] the importance of objective measures was reinforced by demonstrating that observer consensus of cosmetic outcome is difficult to obtain. The same line of action had followers in Limbergen [9], Tsouskas [15] and Vrieling [16], among others.

Noguchi [11] measured the breasts' asymmetry objectively with a Moire topography camera. Breast atrophy, skin change, and surgical scar were assessed subjectively by observers. The overall cosmetic outcome was the sum of the individual scores of the objective and subjective assessments, this way introducing the roots for an overall objective assessment.

Al-Ghazal [10,17] correlates the overall subjective evaluation performed by a six-member panel and the patients themselves, with an overall ''objective'' assessment, taking into account the objective measures of breast retraction and nipple deviation, and subjective factors (skin atrophy, skin changes, such as telangectasia or oedema, and surgical scar), analogously to Noguchi [11]. Yet using the same reasoning, Krishnan [18] added four individual rating (difference in volume, breast asymmetry, fibrosis and telangiectasia) to create an overall cosmetic index.

From this quick snapshot of what has been proposed so far, it is easy to conclude that current objective assessment evaluation methods lack a general and consistent approach. Our work aims at a totally objective overall measure, based on a more principled approach, rather that just the sum of the individual indices.

**Figure 1** Positions used in the photographs.

## 2. Data

### 2.1. Study population

This work adopted part of data collected in three different institutions in Portugal, comprising data from 120 patients. All patients were treated with conservative breast surgery, with or without auxiliary surgery, and whole breast radiotherapy, with treatment completed at least one year before the onset of the study. All patients signed an informed consent to participate.

Breast images were acquired employing a 4 M pixel digital camera. Photographs were taken in four positions with the patient standing on floor marks: facing, arms down; facing, arms up; operated side, arms up; contralateral side, arms up. Fig. 1 presents a typical set. A mark was made on the skin at the suprasternal notch and at the midline 25 cm below the first mark. These two marks create a correspondence between pixels measured on the digital photograph and the length in centimetres on the patient.

### 2.2. Reference classification

In order to investigate the possibility of defining a method of assessment reproducible on a worldwide basis, making use of objective measures, a set of patients with known overall classification was required. Since ideally the overall aesthetic assessment should correlate coherently with experts' assessment, collecting this type of evaluation from different areas of the world would probably provide the desired reference classifications. This choice was corroborated by a previous study [8] were it was shown that a homogeneous group of observers with experience in BCCT will provide better inter-observer agreement than a mixed group involving clinicians with different levels of expertise.

Twenty-four clinicians working in 13 different countries were selected, based on their experience in BCCT (number of cases seen per year and/or participation in published work on evaluation of aesthetic results). They were asked to evaluate individually a series of $60 \times 4$ photographs taken from 60 women submitted to BCCT (surgery and radiotherapy). Participants were asked to evaluate overall aesthetic results, classifying each case into one of four categories: *excellent*: treated breast nearly identical to untreated breast; *good*: treated breast slightly different from untreated; *fair*: treated breast clearly different from untreated but not seriously distorted; *poor*: treated breast seriously distorted [1].

In order to obtain a consensus among observers, the Delphi process was used [19,20]. The method is an attempt to obtain expert opinion in a systematic manner. Experts are recruited individually and anonymously. The survey is conducted over several rounds, and the results are analysed and then reported to the group. The process is considered complete when there is a convergence of opinion or when a point of diminishing returns is reached. Consensus was reached in 59/60 cases (98%) after two rounds. The evaluation of each case was considered consensual when more than 50% of observers provided the same classification on aesthetic result.

In order to obtain a larger sample size, a second consensus panel with another 60 cases was initiated. Participants were limited to those who had obtained at least 35 (60%) coincident answers with the final consensus in the previous panel. Nine of the 11 invited experts participated on this second Delphi consensus. Evaluation of each case was considered consensual when more than 67% of observers provided the same classification on aesthetic result. Consensus was obtained in 54/60 cases (88%)—see [21,22] for more details.

Consensus on aesthetic results of breast cancer conservative treatment was obtained by the two panels in 113 of the 120 evaluated patients (94%).

**Table 1** Distribution of the 113 patients over the four classes

| Class | # cases |
|---|---|
| Excellent | 14 |
| Good | 64 |
| Fair | 24 |
| Poor | 11 |
| Total | 113 |

Fourteen patients were classified as *excellent* (12%), 64 as *good* (57%), 24 as *fair* (21%) and 11 as *poor* (10%), as summarized in Table 1.

## 3. Method

To evaluate the aesthetical result of BCCT, an observer identifies and evaluates colour, shape, geometry, irregularity and roughness of the visual appearance of the treated breast, comparing with the untreated breast; then, every aspect is subjectively combined as an overall classification.

Instead of heuristically weighting the individual indices in an overall measure, we introduced pattern classification techniques to find the correct contribution of each individual feature in the final result and the scale intervals for each class, constructing in this way an optimal rule to classify patients.

In order to apply the proposed methodology (a) one must be in possession of a set of patients with known overall cosmetic classifications (as just described); (b) suitable features must be chosen to discriminate classes; finally (c) the optimum separating boundaries between classes must be found.

### 3.1. Feature selection

As possible *types* of features we considered those already identified by domain experts as relevant to the aesthetic evaluation of the surgical procedure [4,9]. It is commonly accepted that the cosmetic result after BCCT is mainly determined by visible skin alterations or changes in breast volume or shape. Skin changes can consist of a disturbing surgical scar or radiation-induced pigmentation or telangiectasia [9].

This information suggests that, to obtain a discriminative, robust, and concise representation of the aesthetical result of BCCT, different *types* of features are to be extracted from the digital images, conveying the perceptual information determining the aesthetical result.

Having selected the *types* of features, here emerges the question as how to extract the features of those types and which type of classifier to use. These issues are discussed at length in the following subsections.

#### 3.1.1. Asymmetry features

There are many ways to describe breast asymmetry. The breast retraction assessment (BRA) is probably the most widely used index. Nevertheless, different measures were progressively introduced in the literature to capture breast asymmetry.

Instead of limiting any subsequent analysis to an initial choice of an asymmetry index, we decided on recording all well-known asymmetry indices — and some new ones introduced in this work — with the purpose of proceeding later to a feature selection analysis. The indices recorded to assess breast asymmetry were the following:

- Breast retraction assessment (BRA) $= \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$: quantifies the difference in nipple position between both breasts and reflects the degree of breast retraction.
- Lower breast contour (LBC): difference between levels of inferior breast contour.
- Upward nipple retraction (UNR) $= |Y_1 - Y_2|$: difference between nipple levels.
- Breast compliance evaluation (BCE) $= |\mathrm{NI}_1 - \mathrm{NI}_2|$: difference between left and right nipple to inframammary fold distance (NI).
- Breast contour difference (BCD): the difference between lengths of left and right breast contours.
- Breast area difference (BAD): the difference between areas of left and right breasts.

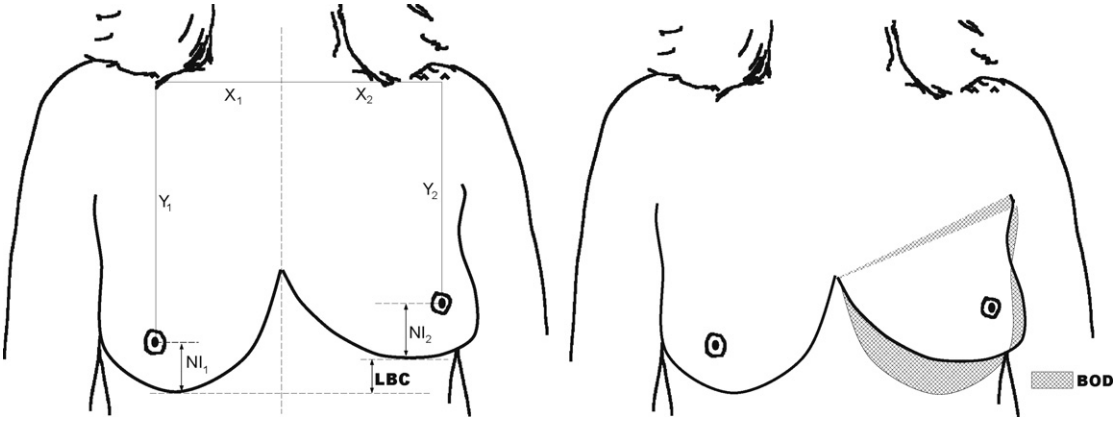


**Figure 2** Illustration showing the lines of measurement.

**Table 2** Dimensionless asymmetry measures

| | |
|---|---|
| pBRA | $\frac{\text{BRA}}{(\sqrt{X_1^2+Y_1^2}+\sqrt{X_2^2+Y_2^2})/2}$ |
| pLBC | $\frac{\text{LBC}}{(Y_1+\text{NI}_1+Y_2+\text{NI}_2)/2}$ |
| pUNR | $\frac{\text{UNR}}{(Y_1+Y_2)/2}$ |
| pBCE | $\frac{\text{BCE}}{(\text{NI}_1+\text{NI}_2)/2}$ |
| pBCD | $\frac{\text{BCD}}{(L_1+L_2)/2}$ |
| pBAD | $\frac{\text{BAD}}{(\text{right area}+\text{left area})/2}$ |
| pBOD | $\frac{\text{BOD}}{(\text{right area}+\text{left area})/2}$ |

- Breast overlap difference (BOD): the non-overlapping area of the two breast after flipping one of them along a vertical line and making coincident both points of junction with thorax (see Fig. 2).

All these indices need a scale to be correctly evaluated—scale provided by the cross markers made on the patient's skin. To investigate the sufficiency of dimensionless features to portrait the overall aesthetic result, a set of corresponding dimensionless features were also defined and recorded from these seven base features (see Table 2).

Note that these dimensionless features do not depend on a scale mark, but only on the incisura jugularis position. The sufficiency of these dimensionless features would simplify the extraction of the measures, as it would render unnecessary the scale mark.

The asymmetry features were extracted with the help of a set of key points signalled manually by the user in the image, using the BCCT.core software, developed specifically for this purpose. The sternal notch, the scale mark, and the two nipples were identified with the help of moving markers. Breast contours were adjusted with an active contour based on splines with 11 control points.

### 3.1.2. Colour difference features

To mitigate the influence of variation of the image capturing conditions on image colour, a correction operation was applied before the extraction of any colour based feature.

Histogram equalization is a well-known image contrast enhancement method. However, the pure equalization forces the output to have a uniform pixel distribution, often leading to unnatural effects and visually disturbing artefacts. We opted for the method proposed in [23] that constructs an output histogram as a "weighted mean" of the input histogram and the perfectly uniform histogram. Setting the trade-off parameter $\alpha$ to 0 one obtains the original image unchanged, setting $\alpha$ to 1 the histogram is perfectly equalized. For colour images the histogram processing can be applied on the red, green and blue channels independently. However, that may change the order relation of the RGB components and thus producing hue-shifting related artefacts. Instead of going for the complex solution proposed in [23]— with unacceptable waiting times for the user when deployed for real operation — we opted to work with the data from the three channels together, as if providing from a single extended channel. That is enough to preserve the relative order of the RGB components.

To avoid distortions on the histogram correction due to different backgrounds, this histogram equalization was preceded by foreground/background segmentation, applying the histogram correction to the patient area only. For this purpose the foreground/background segmentation does not need to be very precise. A simple colour based segmentation, as detailed in the annex, suffices. Fig. 3 presents an example of an image before and after the histogram preprocessing.

After preprocessing the image by histogram equalization, the colour space was converted from

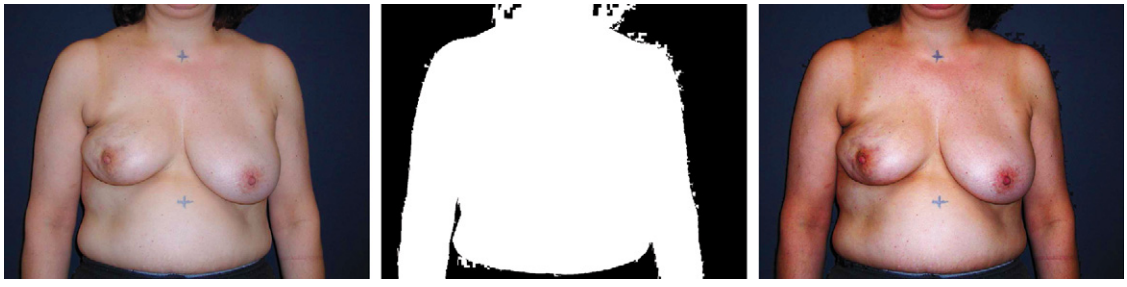

**Figure 3** An image before (left) and after (right) the histogram equalization ($\alpha = 0.6$). The patient mask is shown in the middle position.

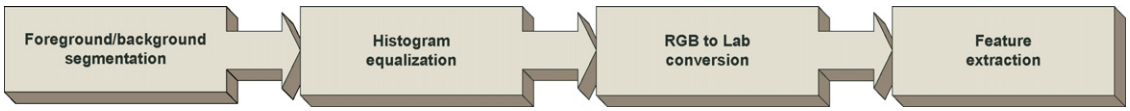


**Figure 4** Chain of operations for the extraction of colour features.

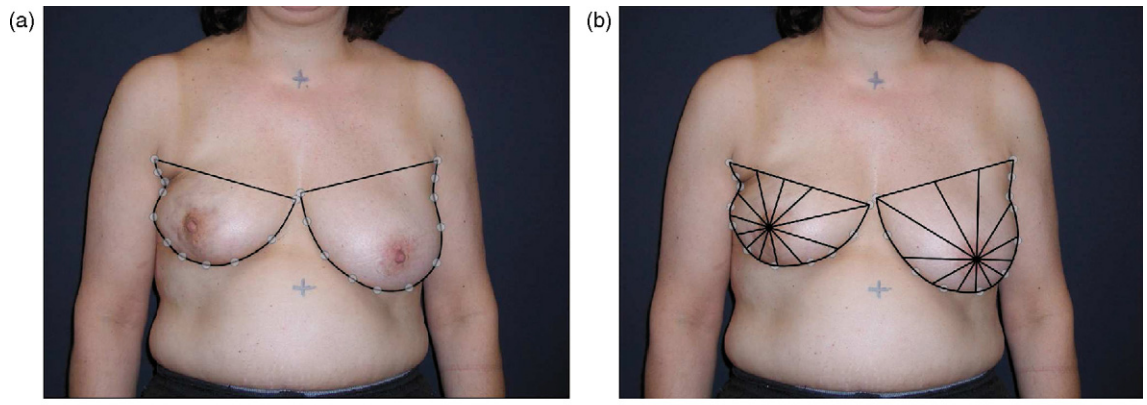


**Figure 5** Masks for histogram based measures.

the RGB values typically obtained from most image acquisition equipment to a more perceptually uniform colour space. The adopted colour space was the CIE $L^*a^*b^*$. All colour features were extracted after these preprocessing operations, as depicted in the block diagram of Fig. 4.

Measuring the dissimilarity between (parts of) images is of central importance in a broad range of applications, with a natural coexistence of different measures [24]. The radiation-induced pigmentation has a global effect over the colour of the treated breast. Therefore, we recorded two global measures based on histograms. First, breasts were selected with the help of two region-masks, as illustrated in Fig. 5 (a). Next, for each breast the 3D colour histogram was computed, consisting of $N = 512$ ($8 \times 8 \times 8$) bins. We also recorded the marginal histograms for each of the three channels, but with $M = 64$ bins each. For each of these four histograms ($L$, $a$, $b$, and Lab3D), two dissimilarity measures were computed, the $\chi^2$ statistic and the earth movers distance (EMD) [24]. That amounts to eight different indices portraying global colour dissimilarity, hereafter represented as (the c prefix stands for colour)

$$c\chi^2_L, c\chi^2_a, c\chi^2_b, c\chi^2_{\text{Lab3D}}, \text{cEMD}_L, \text{cEMD}_a, \text{cEMD}_b, \text{cEMD}_{\text{Lab3D}}.$$

### 3.1.3. Scar visibility features

The surgical scar apperance was translated on a localized colour difference. To compute a local colour difference, the following procedure was adopted:

(1) each breast was divided into 12 angular sectors (30°), with vertice on the nipple, see Fig. 5(b).
(2) the colour histogram for each sector was computed ($L$, $a$, $b$, Lab3D).
(3) the similarity between corresponding sectors was computed as for the global colour change, using $\chi^2$ statistic and EMD.
(4) the maximum value of each pair of corresponding sectors was recorded. That amounts for eight different values.

## 3.2. Pattern classifier

Different alternatives are at the disposal of the experimenter to design classifiers for multiclass problems. However, the problem here addressed, as many real life problems, involves classifying examples into classes which have a natural ordering. Conventional methods for nominal classes or for regression problems could be employed to solve ordinal data problems. However, the use of techniques designed specifically for ordered classes yields simpler classifiers, making it easier to interpret the factors that are being used to discriminate among classes. The reduced number of examples available also suggested the adoption of techniques with known good performance under these conditions, such as support vectors machines (SVMs) [25]. Our previous results [26,27] support this decision.

### 3.2.1. SVM basics

We consider briefly how the SVM binary classifier is formulated. In a typical binary classification problem we have a (training) set of points from two different classes and we are interested in separating them by a hyper plane. This is a typical form of a linear classifier. There are many linear classifiers that might separate perfectly the points of the two classes. To select the surface best suited to the task, the SVM sets the linear decision function in the middle of the gap between classes, correctly classifying all the training samples and putting the decision function as far apart from the given samples as possible: this way we minimize the chance of making an incorrect prediction over the unseen examples.

In the cases where the classes are not linearly separable in the feature space, we can no longer find a hyper plane perfectly separating the points of the two classes. If the non-linearly separable data can be interpreted as the result of noisy points (measurement errors, uncertainty in class membership, etc.), we can still keep the linear classifier and accept some errors. The goal is now to make the margin as large as possible but at the same time to keep the number of points misclassified as small as possible. Therefore the computation of the hyper plane becomes more involved, with the need to control the trade-off between the dual objectives of maximizing the margin and minimizing the misclassification error. A penalty for misclassifying an example is added to the objective function, weighted by a parameter $C$ (for the standard C-SVM formulation, as used in this work), which controls the trade-off between misclassifying the data and the achieved margin. A high $C$ value will force the SVM training to avoid classification errors.

If the non-linearly separable data portraits some intrinsic property of the problem, a more complex classifier, allowing more general boundaries between classes, may be more appropriate. The general methodology is to map the input feature space to a high dimensional feature space where the classes can be satisfactorily separated by a hyper plane. Then, the (linear) SVM method can be mobilized for the design of the hyper plane classifier in the new feature space. However, there is an elegant property in the SVM methodology allowing the implicit mapping into high dimension spaces [28]. This method is based on the idea of kernel functions $k(\mathbf{x}, \mathbf{y})$. Using a kernel function, one can impose nonlinear boundaries, without explicitly knowing the mapping to the high dimension space. Although the SVM formulation does not include criteria to select a kernel function that gives good generalisation (or results in a classifier with low expected error bound), some common kernels include polynomial and radial basis functions $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2)$.

### 3.2.2. Classification of ordinal data

Frank and Hall [29] introduced a simple algorithm that enables standard classification algorithms to exploit the ordering information in ordinal prediction problems. First, the data is transformed from a $K$-class ordinal problem to $K-1$ binary class problems. Training of the $i$th classifier is performed by converting the ordinal dataset with classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ into a binary dataset, discriminating $\mathcal{C}_1, \ldots, \mathcal{C}_i$ against $\mathcal{C}_{i+1}, \ldots, \mathcal{C}_K$; in fact it represents the test $\mathcal{C}_{\mathbf{x}} > i$. To predict the class value of an unseen instance, the $(K-1)$ binary outputs are combined to produce a single estimation. Any binary classifier can be used as the building block of this scheme, in particular a binary SVM classifier.

Observe that, under Frank and Hall approach, each of the $(K-1)$ boundaries is computed independently. That may result in intersecting boundaries. As the intersection point of two boundaries would indicate an example with three or more classes equally probable (not plausible with ordinal classes), we introduced in [26] a method, *the data replication method*, that restricts the search to non-crossing boundaries.

The Frank and Hall method and the data replication method, both instantiated with SVMs, were used to train a classifier in the feature space previously introduced.

### 3.2.3. Classification power of features

It is obvious that neither colour nor asymmetry or scar visibility features alone are able to provide low classification error. We would like to select the subset of features attaining the lowest classification error. Evaluating all possible subsets is unfeasible. Remember that we recorded $7+7+4+4+4+4=30$ features, leading to $2^{30}$ different subsets of features. Therefore, we performed an educated partial full-search. Call

(1) $\mathcal{A}_1$ the set of asymmetry features with dimension. $\mathcal{A}_1 = \{\mathrm{BRA}, \mathrm{LBC}, \mathrm{UNR}, \mathrm{BCE}, \mathrm{BCD}, \mathrm{BAD}, \mathrm{BOD}\}$
(2) $\mathcal{A}_2$ the set of dimensionless asymmetry features. $\mathcal{A}_2 = \{\mathrm{pBRA}, \mathrm{pLBC}, \mathrm{pUNR}, \mathrm{pBCE}, \mathrm{pBCD}, \mathrm{pBAD}, \mathrm{pBOD}\}$
(3) $\mathcal{C}_1$ the set of colour features based of the $\chi^2$ statistic. $\mathcal{C}_1 = \{c\chi^2_L, c\chi^2_a, c\chi^2_b, c\chi^2_{\mathrm{Lab3D}}\}$
(4) $\mathcal{C}_2$ the set of colour features based of the EMD. $\mathcal{C}_2 = \{\mathrm{cEMD}_L, \mathrm{cEMD}_a, \mathrm{cEMD}_b, \mathrm{cEMD}_{\mathrm{Lab3D}}\}$
(5) $\mathcal{S}_1$ the set of scar features based of the $\chi^2$ statistic. $\mathcal{S}_1 = \{s\chi^2_L, s\chi^2_a, s\chi^2_b, s\chi^2_{\mathrm{Lab3D}}\}$

(6) $\mathcal{S}_2$ the set of scar features based of the EMD. $\mathcal{S}_2 = \{\text{sEMD}_L, \text{sEMD}_a, \text{sEMD}_b, \text{sEMD}_{\text{Lab3D}}\}$

Then we considered all possible subsets of features in one of the following forms:

$$\begin{aligned}
&\text{subset}(\mathcal{A}_1) \cup \text{subset}(\mathcal{C}_1) \cup \text{subset}(\mathcal{S}_1)\\
&\text{subset}(\mathcal{A}_1) \cup \text{subset}(\mathcal{C}_1) \cup \text{subset}(\mathcal{S}_2)\\
&\text{subset}(\mathcal{A}_1) \cup \text{subset}(\mathcal{C}_2) \cup \text{subset}(\mathcal{S}_1)\\
&\text{subset}(\mathcal{A}_1) \cup \text{subset}(\mathcal{C}_2) \cup \text{subset}(\mathcal{S}_2)\\
&\text{subset}(\mathcal{A}_2) \cup \text{subset}(\mathcal{C}_1) \cup \text{subset}(\mathcal{S}_1)\\
&\text{subset}(\mathcal{A}_2) \cup \text{subset}(\mathcal{C}_1) \cup \text{subset}(\mathcal{S}_2)\\
&\text{subset}(\mathcal{A}_2) \cup \text{subset}(\mathcal{C}_2) \cup \text{subset}(\mathcal{S}_1)\\
&\text{subset}(\mathcal{A}_2) \cup \text{subset}(\mathcal{C}_2) \cup \text{subset}(\mathcal{S}_2)
\end{aligned}$$

We limited further each individual subset to 0, 1 or 2 elements, resulting in 25,137 different subsets of features being evaluated. Due to the shortage of labelled data, the discrimination power of each subset of features was estimated using a five-fold cross-validation scheme [30].

In supervised classification problems with ordered classes, it is common to assess the performance of the classifier using measures which are not really appropriate. Very often, every misclassification is considered equally costly and the misclassification error rate (MER) is used. However, for ordered classes, losses that increase with the absolute difference between the classes numbers are more natural choices in the absence of better information. The mean absolute error deviation (MAD) criterion takes into account the degree of misclassification and thus is richer criterion than MER.

Still, this measure depends on the number assigned to each class, which is somewhat arbitrary. In order to avoid the influence of the numbers chosen to represent the classes on the performance assessment, we should only look at the order relation between "true" and "predicted" class numbers. The use of Kendall's tau-b, $\tau_b$, [31] is a step forward in that direction.

To define $\tau_b$, start with the $N$ data points $(\mathcal{C}_{\mathbf{x}_i}, \hat{\mathcal{C}}_{\mathbf{x}_i}), i = 1, \ldots, N$, associated with the true and predicted classes, and consider all $N(N-1)/2$ pairs of data points. Following the notation in [31], we call a pair $(i, j)$*concordant* if the relative ordering of the true classes $\mathcal{C}_{\mathbf{x}_i}$ and $\mathcal{C}_{\mathbf{x}_j}$ is the same as the relative ordering of the predicted classes $\hat{\mathcal{C}}_{\mathbf{x}_i}$ and $\hat{\mathcal{C}}_{\mathbf{x}_j}$. We call a pair *discordant* if the relative ordering of the true classes is opposite from the relative ordering of the predicted classes. If there is a tie in either the true or predicted classes, then we do not call the pair either concordant or discordant. If the tie is in the true classes, we will call the pair an "extra true pair", $e_t$. If the tie is in the predicted classes, we will call the pair an "extra predicted pair", $e_p$. If the tie is both on the true and the predicted classes, we ignore the pair. The $\tau_b$ coefficient can be computed as

$$\tau_b = \frac{\text{concordant} - \text{discordant}}{\sqrt{\text{concordant} + \text{discordant} + e_t}\ \sqrt{\text{concordant} + \text{discordant} + e_p}}$$

The $\tau_b$ coefficient attains its highest value, 1, when both sequences agree completely, and 0 when the two sequences totally disagree.

**Table 3** Confusion matrix for the best panel expert

| Consensus/expert | Excellent | Good | Fair | Poor | Total |
|---|---|---|---|---|---|
| Excellent | 10 | 4 | 0 | 0 | 14 |
| Good | 10 | 52 | 2 | 0 | 64 |
| Fair | 0 | 5 | 19 | 0 | 24 |
| Poor | 0 | 0 | 4 | 7 | 11 |
| Total | 20 | 61 | 25 | 7 | 113 |

Differences in 22.12% of the cases.

## 4. Results

The analysis of the confusion matrix for the best expert of the panel (the expert with best agreement with the consensus) in Table 3, supported by a preliminary study, revealed that most of the difficulties to correctly estimate the class of a patient would show up in the discrimination between *excellent* and *good* classes. Therefore, we decided to start with three classes our attempt to objectify the aesthetical evaluation, by grouping together the *excellent* and *good* cases in a single class.

### 4.1. Results for three classes

The test results concerning the MAD and $\tau_b$ coefficients are summarized in Table 4 for the first ranked feature subsets, with decreasing accuracy. In parentheses, the standard deviation of the index is provided. The best parameterization of the classifier for each feature subset was found based on the cross-validation scheme. The experiments were conducted for the Frank and Hall method and the data replication method, both with linear and radial basis kernels. We performed a "grid-search" on the parameter $C$ using cross-validation. We tried exponentially growing sequences of $C$: $C = 1.25^{-1}, 1.25^{0}, 1.25^{1}, \ldots, 1.25^{30}$. Experiments were conducted with features extracted from images of patients in front position, both with arms down and up. Best results were attained with the data replication method using the radial basis kernel (gamma parameter set to 3, empirically tuned for

**Table 4** Mean (standard deviation) of $\tau_b$ and MAD over the best-ranked subset of features

| Feature subset | $\tau_b$ | MAD |
|---|---|---|
| $\{pLBC, pBCE, cEMD_a, s\chi^2_{Lab3D}\}$ | 0.79 (0.084) | 0.15 (0.050) |
| $\{pLBC, pBCE, cEMD_a, cEMD_{Lab3D}\}$ | 0.77 (0.138) | 0.14 (0.071) |
| $\{pLBC, pBCE, cEMD_L, cEMD_a\}$ | 0.77 (0.138) | 0.14 (0.071) |
| $\{pLBC, pBCE, cEMD_a, s\chi^2_L\}$ | 0.77 (0.130) | 0.17 (0.057) |
| $\{pLBC, pBCE, cEMD_a, cEMD_{Lab3D}, s\chi^2_a\}$ | 0.76 (0.099) | 0.16 (0.066) |

**Table 5** Confusion matrix for the three class experiment

| Consensus/prediction | Excellent/good | Fair | Poor | Total |
|---|---|---|---|---|
| (a) Feature subset $\{pLBC, pBCE, cEMD_a, s\chi^2_{Lab3D}\}$. Differences in 12.39% of the cases | | | | |
| Excellent/good | 75 | 3 | 0 | 78 |
| Fair | 3 | 20 | 1 | 24 |
| Poor | 1 | 6 | 4 | 11 |
| Total | 79 | 29 | 5 | 113 |
| **Consensus/expert** | **Excellent/good** | **Fair** | **Poor** | **Total** |
| (b) Best panel expert result. Differences in 9.73% of the cases | | | | |
| Excellent/good | 76 | 2 | 0 | 78 |
| Fair | 5 | 19 | 0 | 24 |
| Poor | 0 | 4 | 7 | 11 |
| Total | 81 | 25 | 7 | 113 |

best performance), for the arms down position, for which results are presented.

The main assertion concerns the superiority of the LBC and BCE, and corresponding dimensionless versions, over the other asymmetry indices considered in this study. In fact, LBC and BCE are the only asymmetry indices present in the top-10 subsets. It is also clear that the dimensionless features revealed better suitability to the discrimination process than the with-dimension features. These conclusions may embody the greater robustness of LBC and BCE against adverse conditions on the capture process of the images. Patients may not adopt the correct posture; the photographer may take the image not exactly in the front position, etc. The happy fact of dimensionless features exhibiting better performance than the corresponding dimension-based features, simplifies the capture process, dismissing the need of auxiliary marks.

**Figure 6** Patient that was erroneously classified by the automatic algorithm to non-contiguous classes. Patient #103. Consensus: Poor. Automatic: Excellent or Good.

Table 5 presents the confusion matrices for the automatic algorithm based on the feature subset $\{pLBC, pBCE, cEMD_a, s\chi^2_{Lab3D}\}$ and for the panel expert most approximating the reference classification (consensus). As seen, the performance of the automatic algorithm is comparable with the performance of the best panel expert in the set of 113 patients. The expected absolute error rate of 15.7% was considered acceptable.

In Fig. 6 is presented the example that was erroneous classified by the automatic algorithm to a non-contiguous class. In patient #103 the quite visible deformation of the treated breast is not reflected in a significative asymmetry on the nipple position or the lower breast contour.

## 4.2. Results for four classes

Following the procedure adopted for the three classes setup, we evaluated the classifiers and the

**Table 6** Confusion matrix for automatic algorithm

| Consensus/ prediction | Excellent | Good | Fair | Poor | Total |
|---|---|---|---|---|---|
| Excellent | 0 | 14 | 0 | 0 | 14 |
| Good | 0 | 61 | 3 | 0 | 64 |
| Fair | 0 | 3 | 21 | 0 | 24 |
| Poor | 0 | 1 | 7 | 3 | 11 |
| Total | 0 | 79 | 31 | 3 | 113 |

Differences in 16.81% of the cases.

**Table 7** Confusion matrix for automatic algorithm after reassignment of misclassification costs

| Consensus /prediction | Excellent | Good | Fair | Poor | Total |
|---|---|---|---|---|---|
| Excellent | 12 | 2 | 0 | 0 | 14 |
| Good | 8 | 53 | 3 | 0 | 64 |
| Fair | 1 | 2 | 20 | 1 | 24 |
| Poor | 0 | 0 | 1 | 10 | 11 |
| Total | 21 | 57 | 24 | 11 | 113 |

Differences in 15.9% of the cases.

feature subsets at our disposal. The best result was again for the subset $\{\text{pLBC}, \text{pBCE}, \text{cEMD}_a, s\chi^2_{\text{Lab3D}}\}$, with the data replication method using a radial base kernel. The accuracy attained was of MAD = 0.296 (0.019) and $\tau_b = 0.697$ (0.107). The confusion matrix for 113 patient dataset is presented in Table 6.

No patients were classified as *Excellent* by the algorithm. This is mainly due to the unbalancing of the data, with the *Good* class being much more represented than the *Excellent* class. Moreover, humans tend to use more the middle classes in detriment of extreme classes. These facts led us to adopt higher costs to the misclassifications of the cases of the Excellent and Poor classes. When training a classifier, the simplest setting is to equally penalize any misclassification error. (In the C-SVM formulation higher $C$ values mean higher penalties.) However, more general settings allow us to apply different cost values to examples from different classes. The classifiers were then trained with the costs ($\text{Cost}_{\text{Exc}}$, $\text{Cost}_{\text{Good}}$, $\text{Cost}_{\text{Fair}}$, $\text{Cost}_{\text{Poor}}$) = (2.5; 1; 1; 2). Under this formulation the best performance was obtained with the classifier of Frank and Hall based on the feature subset $\{\text{pLBC}, \text{pBCE}, \text{cEMD}_L, \text{cEMD}_a, \text{cEMD}_b, s\chi^2_{\text{Lab3D}}\}$, trained with a radial basis function (RBF) kernel, with $C = 5.9605, \gamma = 2$. The final weighted classifier exhibits an expected MAD of MAD = 0.348 (0.141) and a kandall coefficient of $\tau_b = 0.656$ (0.175).

The confusion matrix, presented in Table 7, reveals a more appropriate distribution of cases amongst classes, comparable with the confusion matrix for the best expert. As observed, just one patient was erroneously classified to a non-contiguous class. Patient #049 was classified automatically as *excellent* while being classified as *fair* by the panel (see Fig. 7).

**Figure 7** Patient erroneously classified by the automatic algorithm to a non-contiguous classes. Patient #049. Consensus: Fair. Automatic: Excellent.

## 5. Conclusion

To evaluate the aesthetical result of BCCT, an observer identifies and evaluates colour, shape, geometry, irregularity and roughness of the visual appearance of the treated breast, comparing with the untreated breast. This type of examination is rather subjective, individual and to a great extent depends on the observer's experience. An ability to obtain objective measures of these features would be very helpful for assuring objective analysis of the aesthetical result of BCCT, and creating systematic databases for education, health care, and research purposes. The surge of data mining techniques has created new exciting possibilities in medicine.

In this paper, we have presented an approach to an automated analysis of images of BCCT, aiming to categorize the images into *excellent*, *good*, *fair*, and *poor* classes. To accomplish the categorization, first, a concise representation of a BCCT image is obtained based on asymmetry, colour, scar visibility features. The representation is then further analyzed by a pattern classifier performing the categorization.

The asymmetry between both breasts was conveyed under a large set of indices, some of them introduced for the first time in this work. The consideration of two versions of each index, with and without dimension, tried to evaluate the need to have an auxiliary mark, providing the scale correction. The

fact that the best subset of features included only relative indices, dismissed such a necessity.

To extract the colour features, the colour content of each breast is first characterized by a colour histogram; the dissimilarity between both histograms serves as the colour features. The scar visibility was translated into local colour dissimilarity, by comparing corresponding sectors of the breasts. Beyond a doubt a larger number of salient scar features can be extracted. This issue will be the subject-matter of further research.

The data replication method instantiated with support vector machines, a method specific for ordinal classes, provided the highest correct classification rate, outperforming the Frank and Hall method. A correct classification rate of about 70% was obtained when categorizing a set of unseen images into the aforementioned four classes. This accuracy is comparable with the result attained by the best panel expert in the set of the 113 patients. Bearing in mind the high similarity of the decision classes, the results obtained are rather encouraging and the developed tools could be very helpful for assuring objective analysis of the images of breast cancer conservative treatment.

## Appendix A. Software

The BCCT.core Workstation allows the user to select one of the available medical images from the image database. This database may include from one to four images of the same patient, comprising front views of the patient (arms up and down) and lateral views. The selected image(s) is (are) displayed on the monitor screen of the user computer.

The BCCT.core Workstation integrates a tool to quickly and accurately automate the measurement process of all the well-known indices correlated with the overall aesthetical result of the breast cancer conservative treatment. Besides the indices describing asymmetry between breasts, the colour difference and surgical scar visibility were also conveyed with appropriate indices. All these measures are reported automatically to the user and are saved in a database. BCCT.core is also especially useful for reporting quickly and accurately results that were, up until now, time consuming, requiring the proficiency on several tools.

A second major functionality of BCCT.core Workstation is the ability to convert automatically the set of measures performed on the digital images onto an overall objective classification of the aesthetical result. During the development phase of the

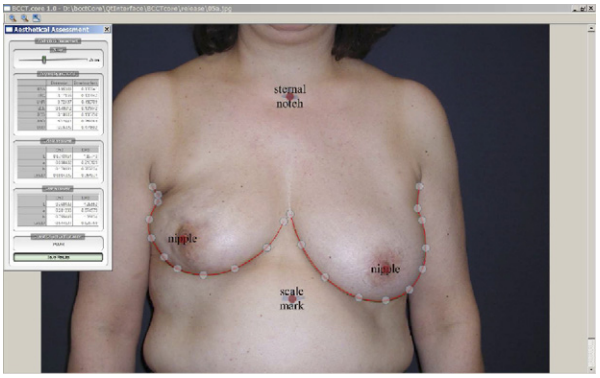


**Figure A.1** Graphical user interface of the BCCT.core software.

BCCT.core the artificial intelligence module described in this article was trained and optimized to predict the overall aesthetical classification on the universally used scale of four classes (*excellent*, *good*, *fair*, *poor*). when in use, the artificial intelligence module automatically evaluates the aesthetical result of the breast cancer conservative treatment. This classification is also reported automatically to the user and saved in a database. The objective overall classification outputted by this module constitutes a valuable summary of the aesthetical result, enabling effective comparison among different medical teams and centres all over the world.

The next figure represents a screenshot of the graphical user interface of the BCCT.core software (Fig. A.1).

## References

[1] Harris JR, Levene MB, Svensson G, Hellman S. Analysis of cosmetic results following primary radiation therapy for stages I and II carcinoma of the breast. Int J Radiation Oncol Biol Phys 1979;5:257—61.

[2] Beadle GF, Silver B, Botnick L, Hellman S, Harris JR. Cosmetic results following primary radiation therapy for early breast cancer. Cancer 1984;54:2911—8.

[3] Pierquin B, Huart J, Raynal M, Otmezguine Y, Calitchi E, Mazeron JJ, et al. Conservative treatment for breast cancer: long-term results (15 years). Radiother Oncol 1991;20:16—23.

[4] Pezner RD, Patterson MP, Hill LR, Vora N, Desai KR, Archambeau JO, et al. Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer. Int J Radiation Oncol Biol Phys 1985;11:575—8.

[5] Sacchini V, Luini A, Tana S, Lozza L, Galimberti V, Merson M, et al. Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer. Eur J Cancer 1991;27:1395—400.

[6] Sneeuw KC, Aaronson NK, Yarnold JR, Broderick M, Ross JRG, Goddard A. Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 1.

comparison of patients' ratings, observers' ratings, and objective assessments. Radiother Oncol 1992;25:153—9.
[7] Christie DRH, O'Brien M-Y, Christie JA, Kron T, Ferguson SA, Hamilton CS, et al. A comparison of methods of cosmetic assessment in breast conservation treatment. Breast 1996;5:358—67.
[8] Cardoso MJ, Santos AC, Cardoso JS, Barros H, Oliveira MC. Choosing observers for evaluation of aesthetic results in breast cancer conservative treatment. Int J Radiation Oncol Biol Phys 2005;61:879—81.
[9] Limbergen EV, Schueren EV, Tongelen KV. Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. Proposal of a quantitative scoring system. Radiother Oncol 1989;16:159—67.
[10] Al-Ghazal SK, Blamey RW, Stewart J, Morgan AL. The cosmetic outcome in early breast cancer treated with breast conservation. Eur J Surg Oncol 1999;25:566—70.
[11] Noguchi M, Saito Y, Mizukami Y, Nonomura A, Ohta N, Koyasaki N, et al. Breast deformity, its correction, and assessment of breast conserving surgery. Breast Cancer Res Treatment 1991;18:111—8.
[12] Clarke D, Martinez A, Cox RS. Analysis of cosmetic results and complications in patients with stage i and ii breast cancer treated by biopsy and irradiation. Int J Radiation Oncol Biol Phys 1983;9:1807—13.
[13] Beadle GF, Come S, Henderson IC, Silver B, Hellman S, Harris JR. The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. Int J Radiation Oncol Biol Phys 1984;10: 2131—7.
[14] Pezner RD, Lipsett JA, Vora NL, Desai KR. Limited usefulness of observer-based cosmesis scales employed to evaluate patients treated conservatively for breast cancer. Int J Radiation Oncol Biol Phys 1985;11:1117—9.
[15] Tsouskas LI, Fentiman IS. Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer. Breast Cancer Res Treatment 1990;15:185—90.
[16] Vrieling C, Collette L, Bartelink E, Borger JH, Brenninkmeyer SJ, Horiot JC, et al. Validation of the methods of cosmetic assessment after breast-conserving therapy in the EORTC boost versus no boost trial. Int J Radiation Oncol Biol Phys 1999;45:667—76.
[17] Al-Ghazal SK, Fallowfield L, Blamey RW. Patient evaluation of cosmetic outcome after conserving surgery for treatment of primary breast cancer. Eur J Surg Oncol 1999;25:344—6.
[18] Krishnan L, Stanton AL, Collins CA, Liston VE, Jewell WR. Form or function? Part 2. Objective cosmetic and functional correlates of quality of life in women treated with breast-conserving surgical procedures and radiotherapy. Cancer 2001;91:2282—7.
[19] Jones J, Hunter D. Consensus methods for medical and health services research. Br Med J 1995;311:376—80.
[20] Hasson F, Keeney S, McKenna H. Research guidelines for the delphi survey technique. J Adv Nurs 2000;32: 1008—15.
[21] Cardoso MJ, Cardoso JS, Santos AC, Barros H, Oliveira MC. Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer. Breast 2006;15:52—7.
[22] Cardoso MJ, Cardoso JS, Santos AC, Vrieling C, Christie D, Liljegren G, et al. Factors determining esthetic outcome after breast cancer conservative treatment. Breast J 2007;13(2):140—6.
[23] Duan J, Qiu G. Novel histogram processing for colour image enhancement. In: Proceeding of the third international conference on image and graphics (ICIG'04); 2004.p. 55—8.
[24] Rubner Y, Puzich J, Tomasi C, Buhmann JM. Empirical evaluation of dissimilarity measures for color and texture. Comput Vision Image Understanding 2001;84:25—43.
[25] Vapnik VN. Statistical learning theory, ser. adaptive and learning systems for signal processing, communications, and control. New York: Wiley; 1998.
[26] Cardoso JS, da Costa JFP, Cardoso MJ. Modelling ordinal relations with SVMs: an application to objective aesthetic evaluation of breast cancer conservative treatment. Neural Netw 2005;18:808—17.
[27] Vapnik VN. SVMs applied to objective aesthetic evaluation of conservative breast cancer treatment. In: Proceedings of the international joint conference on neural networks (IJCNN) 2005, vol. 4; 2005. p. 2481—6.
[28] Theodoridis S, Koutroumbas K. Pattern recognition. San Diego: Academic Press; 2003.
[29] Frank E, Hall M. A simple approach to ordinal classification. In: Proceedings of the 12th European conference on machine learning, vol. 1; 2001. p. 145—56.
[30] Hastie RT, Friedman RTJ. The elements of statistical learning: data mining, inference, and prediction, ser. Springer series in statistics. New York: Springer; 2001, August.
[31] Press W, Flannery B, Teukolsky S, Vetterling W. Numerical recipes in C: the art of scientific computing. Cambridge: Cambridge University Press; 2002.