SVMs Applied to Objective Aesthetic Evaluation of Conservative Breast Cancer Treatment

Jaime S. Cardoso
Faculdade de Engenharia
Universidade do Porto / INESC Porto
Portugal
E-mail: jaime.cardoso@ieee.org

Joaquim F. Pinto da Costa
Faculdade de Ciências
Universidade do Porto
Portugal
E-mail: jpcosta@fc.up.pt

Maria J. Cardoso
Faculdade de Medicina
Universidade do Porto
Portugal
E-mail: mjcard@med.up.pt

***Abstract*— Cosmetic assessment of conservative breast cancer treatment plays a major role in the study of breast cancer techniques. Objective assessment methods are being preferred to overcome the drawbacks of subjective evaluation. In this paper a methodology for the objective assessment of conservative breast cancer treatment is proposed. The quantitative measures used in this research provide an objective way to calculate the overall cosmetic result. We report experiments using support vector machines to derive an optimal assessment rule. The results seem to indicate that it is possible to construct an algorithm for a complete objective classification of the aesthetic result of breast conservative treatment.**

I. Introduction

Breast cancer treatment has evolved considerably in the last decade, with the widespread adoption of breast-conserving approaches to manage early breast cancer. Breast-conserving approaches aim not only at local tumour control and survival rates equivalent to mastectomy but also at better aesthetic results. Although considerable research has been put on breast-conserving techniques, different forms of performing them, as well as incorrect working practices, contribute to different aesthetic results.

Traditionally, cosmetic assessment has been performed subjectively by a group of observers. However, this form of assessment is difficult to reproduce and dismisses any possible comparison. The assessment is often not easily performed and the interpretation of the aesthetic result can depend heavily on the skill of the observer. Observers with different background evaluate differently, M. J. Cardoso [1]. This has drawn more attention to objective evaluation.

Early work to overcome the drawbacks of subjective assessment has included the use of quantitative measures. However, these efforts have merely correlated objective values with the subjective overall assessment, without attempting to conclude the overall aesthetic result solely from objective features.

Having introduced the problem, we will next review previous related work, paving the way for the second section, where we present a novel method of assessment, based on pattern classification tools. Section III briefly describes the details and the results of the experiments carried on. We discuss the problem of ranking patients with the use of a classifier for ordinal categorical data. Section IV concludes the paper with discussion on ongoing and future works.

Brief history of the cosmetic assessment measures

Harris [2] 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 [3], the reported works – of which [4] and [5] 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 [3] the objective assessment of the cosmetic result of the surgery was introduced with the first objective measure to evaluate one of the aspects of cosmesis: Breast Retraction Assessment, BRA. In Pezner [6] 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 [7], Tsouskas [8] and Vrieling [9], among others.

Noguchi [10] 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.

In [11], [12] the authors correlate the overall subjective evaluation performed by a 6-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 [10]. Yet using the same reasoning, Krishnan [13] summed 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 principle approach, rather that just the sum of the individual indices.

0-7803-9048-2/05/$20.00 ©2005 IEEE

II. Data and Method

Instead of heuristically weight 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; *b)* suitable features must be chosen to discriminate classes; and finally *c)* the optimum separating boundaries between classes must be found.

A. 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 expert opinions from different areas of the world would probably provide the desired reference classifications.

Using a Delphi methodology [14], [15], a consensus overall evaluation was collected from some of the most prominent experts in the field, as detailed in M. J. Cardoso [16], for a set of 60 patients. This provided a set of patients with a *reference classification* (also known as *gold standard* or *ground truth*) to reproduce through objective features. Proceeding this way, each and every patient was classified in one of the four categories (table I): *poor*, *fair*, *good*, and *excellent*.

Class	# cases
Poor	7
Fair	12
Good	32
Excellent	9

TABLE I
DISTRIBUTION OF PATIENTS OVER THE FOUR CLASSES.

B. Feature selection

The degree of class separability depends strongly on the choice of the descriptors. As such, assessing the discriminative capability of features is an important task, when choosing between alternative feature sets. Simultaneously, and ideally, the overall aesthetic assessment should rely on simple quantitative and objective measures.

As possible objective features we considered those already identified by domain experts as relevant to the aesthetic evaluation of the surgical procedure [3], [8], [7]. In this way, we started with the commonly used Breast Retraction Assessment (BRA), Lower Breast Contour (LBC), and Upward Nipple Retraction (UNR) measures, which were recorded for each patient.

Pictures were taken from 60 women submitted to conservative breast cancer treatment with a digital camera in four positions (front arms up and down, left and right side arms up), figure 1. For each of the 60 cases, BRA, LBC and UNR were calculated using the digital image, both with arms up and down. Measures were taken directly in a computer. BRA was computed as $BRA = \sqrt{(X_r - X_l)^2 + (Y_r - Y_l)^2}$; LBC was formulated as $LBC = abs(L_r - L_l)$ and UNR as $UNR = abs(Y_r - Y_l)$, figure 2.

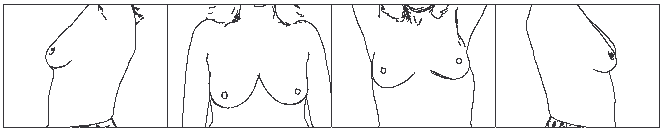


Fig. 1. Positions used in the photographs.

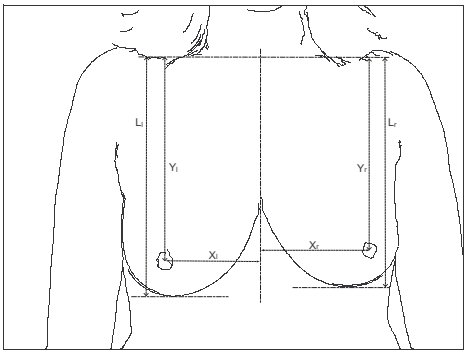
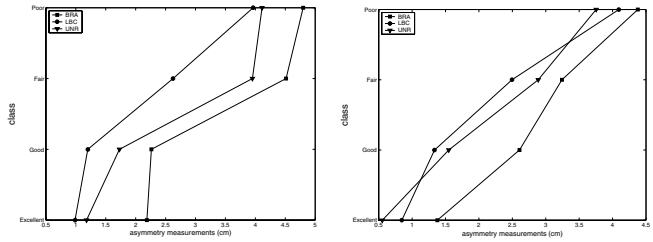


Fig. 2. Illustration showing the lines of measurement.

The measurements collected for BRA, LBC and UNR were correlated with the subjective overall score from the panel of experts. Median values were calculated for each class (intra-class values), both with arms up and down, figure 3 and table II. The dispersion of values inside each class was measured by the standard deviation. As easily concluded visually from figure 3, the three measures correlate well with the panel scoring.



(a) Arms down. (b) Arms up.

Fig. 3. Correlation of asymmetry measurements and cosmetic panel scoring.

Although the median values alone suggest that measurements made with arms up provide a better indication of the overall subjective score, the larger standard deviation of measurements in this position, implies otherwise. In fact, the relatively large values of dispersion may indicate that breast asymmetry (as measured by BRA, LBC, UNR or similar

TABLE II

INTRA CLASS VALUES

(a) LBC

Class	LBC, arms down		LBC, arms up	
	median	std dev	median	std dev
Excellent	0.99	0.50	0.85	0.88
Good	1.21	0.83	1.34	0.91
Fair	2.62	1.11	2.50	1.07
Poor	3.96	0.90	4.09	1.71

(b) BRA

Class	BRA, arms down		BRA, arms up	
	median	std dev	median	std dev
Excellent	2.19	0.80	1.38	1.18
Good	2.29	1.33	2.61	1.23
Fair	4.51	2.04	3.24	1.63
Poor	4.80	1.79	4.38	2.98

(c) UNR

Class	UNR, arms down		UNR, arms up	
	median	std dev	median	std dev
Excellent	1.18	0.67	0.55	0.92
Good	1.73	1.33	1.55	1.23
Fair	3.95	1.83	2.89	1.41
Poor	4.11	1.24	3.75	2.49

measures) is not enough to interpret the overall cosmetic score. Other factors, such as visible skin changes, contribute to the aesthetic result of surgery.

> The cosmetic result after breast conserving treatment is mainly determined by visible skin alterations or changes in breast volume or shape. Skin changes may consist of a disturbing surgical scar, or radiation-induced pigmentation or teleangiectasia. — Limbergen 1989 [7]

The suspicion that BRA is not sufficient to successfully separate the four classes is confirmed by visual inspection, figure 4. There is considerable overlap of the classes as BRA varies from low to high values. The same conclusion is obtained using percentage BRA (pBRA), LBC or percentage LBC (pLBC), UNR or percentage UNR (pUNR). This is expected as these measures carry similar information – they are highly correlated, [7].

> Percentage BRA was defined as $pBRA = BRA/\max(\sqrt{X_r^2+Y_r^2}, \sqrt{X_l^2+Y_l^2})$.
> Percentage LBC was defined as $pLBC = LBC/\max(L_r, L_l)$.
> Percentage UNR was defined as $pUNR = UNR/\max(Y_r, Y_l)$.

In order to improve the discriminative capability of the objective measures, we evaluated pairs of features. The features should be as uncorrelated as possible and portrait complementary information about the aesthetic result. The data is plotted in figure 5 for some pairs of selected features. As observed, the discriminative capability has increased appreciably. Particularly in the (BRA, Scar) scatter plot, it is reasonably easy to draw rough boundaries between classes.

> The lack of objective measures to score the scar visibility and skin colour changes after treatment in the published research, led us to use subjective quantities at this phase of the work. Besides scoring the overall cosmetic result, the experts constituting the Delphi panel [16], were also asked to assess these two individual items, using a four grade scale (0-3). The mean and median of the answers for each patient were used as features in the current model. This procedure allows the assessment of the proposed model, without degrading its performance by incorrect feature estimation. Once the model is firmly established, objective measures will replace the panel scoring.

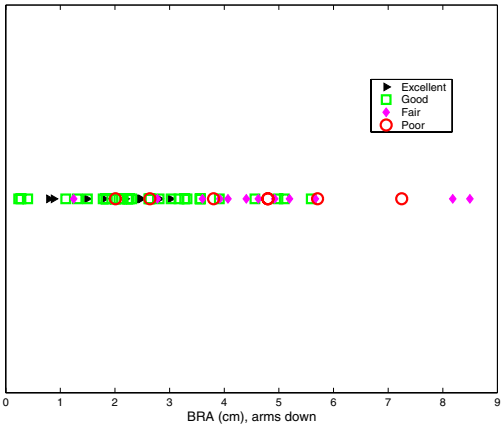
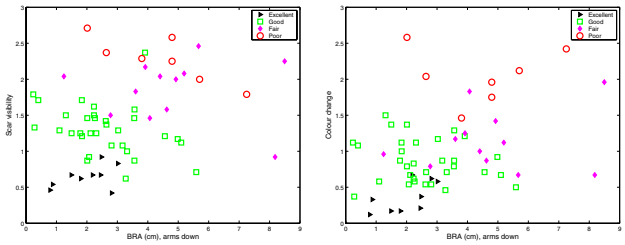
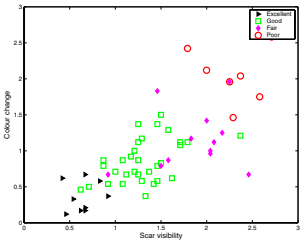


Fig. 4. Variability of BRA over the four classes.



(a) BRA versus Scar visibility. (b) BRA versus colour change.



(c) Scar visibility versus colour change.

Fig. 5. Pairs of features.

To improve even further the quality of the feature set, increasing the class separability, we considered sets of three features: BRA (or similar), scar visibility and skin colour change, figure 6.

The dataset can now be adequately separated: it is possible to establish acceptable decision boundaries between classes.

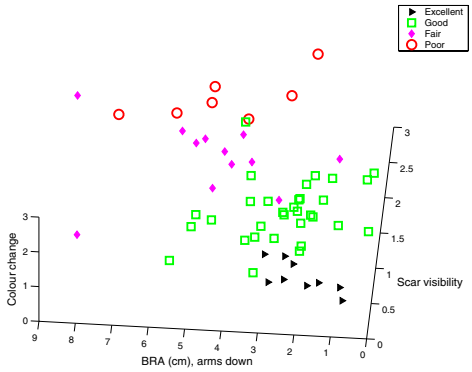


Fig. 6. Three features space.

C. Design of the classifier

As a fast visual checking of the quality of the data shows (figure 6), there is a data value that is logically inconsistent with the others: an individual (patient #31) labeled as *good* when in fact it is placed between *fair* and *poor* in the feature space. Therefore, this individual was not considered in the design of the classifiers (it would not help the learning phase), leaving 59 cases for that task. The *leave one out* method [17] was selected for validation: the classifier is trained 59 times, each time using the available dataset from which a single patient has been deleted; each resulting classifier is then tested on the single deleted patient.

When in possession of a *nearly separable* dataset, a simple linear separator is bound to misclassify some points. But the question that we have to ask ourselves is if the *non-linearly-separable* data portraits some intrinsic property of the problem (in which case a more complex classifier, allowing more general boundaries between classes may be more appropriate) or if it can be interpreted as the result of *noisy points* (measurement errors, uncertainty in class membership, etc), in which case keeping the linear separator and accept some errors is more natural. Supported by Occam's razor principle ("one should not increase, beyond what is necessary, the number of entities required to explain anything"), the latter was the option taken in this research.

In the succeeding section we present, evaluate and compare two different classifiers based on support vector machines and discuss the results achieved with each.

III. Results

The design of the classifier, based on the selected features, followed the large margin principle, setting the boundary as far as part from the available data as possible, minimizing this way the chance of making an incorrect prediction over unseen cases.

Although SVMs have been generalized for multiclass problems [18], in this work, and because the categories are ordered, we performed three binary classifications, (*Poor*/*Fair*, *Fair*/*Good* and *Good*/*Excellent*), which can be interpreted as a simplification of the approach suggested by Hastie and Tibshirani [19]. An issue with this approach is that we are ignoring the ordering information between classes, treating the problem as a nominal classification problem. At most, the ordering is used to reduce the number of binary discriminations to perform. Each pair of classes is separated by a linear boundary, ignoring possible dependencies.

Informally speaking, designing a linear discriminator between two classes translates in choosing a weighted sum of the individual features and the bias to differentiate classes. SVMs try to find the optimal weighted sum and bias. So far we have adopted a different weighted sum for each pair of classes. Suppose for a moment that we were forced to make all boundaries to have the same orientation – to choose the same weighted sum for all decisions. With this additional constraint emerges a monotonic model, where a higher value in an attribute does not lead to a lower decision class. The output of the classifier would be just a set of weights, one for each feature, and a set of biases, the *scale* in the weighted sum. This simplified model captures better the essence of the problem. Another strength of this approach is the reduced number of parameters to estimate, which may lead to a more robust classifier, with greater capacity of generalization. This is the rationale behind the Shashua models [20]. Using the formulation for the fixed margin model [20], we designed a classifier tuned for ordinal categorical data.

To formulate the problem of separating four classes $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ and $\mathcal{C}_4$ using the fixed margin model, consider the training set $\{(x_i, c_i)\}_{i=1}^{N}$, where x_i is the input pattern for the i-th sample and c_i the corresponding desired response, $c_i \in \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$. For each x_i denote

$$y_i^1 = \begin{cases} -1 & \text{if} \quad x_i \in \mathcal{C}_1 \\ +1 & \text{if} \quad x_i \in \mathcal{C}_2 \end{cases} \qquad y_i^2 = \begin{cases} -1 & \text{if} \quad x_i \in \mathcal{C}_2 \\ +1 & \text{if} \quad x_i \in \mathcal{C}_3 \end{cases}$$

$$y_i^3 = \begin{cases} -1 & \text{if} \quad x_i \in \mathcal{C}_2 \\ +1 & \text{if} \quad x_i \in \mathcal{C}_3 \end{cases}$$

Our task can now be summarized as [20]: compute the parameters W, b_1, b_2, b_3 and ξ_j so that to

$$\begin{array}{lll} \text{Minimize} & J(W) \equiv \frac{1}{2}\|W\|^2 + C\sum_j \xi_j & \\ \text{s.t.} & y_i^1(W^T x_i + b_1) \geq 1 - \xi_j & x_i \in \{\mathcal{C}_1, \mathcal{C}_2\} \\ & y_i^2(W^T x_i + b_2) \geq 1 - \xi_j & x_i \in \{\mathcal{C}_2, \mathcal{C}_3\} \\ & y_i^3(W^T x_i + b_3) \geq 1 - \xi_j & x_i \in \{\mathcal{C}_3, \mathcal{C}_4\} \\ & \xi_j \geq 0 & \end{array}$$

We used MATLAB function `quadprog` to solve this quadratic formulation.

Both classifier models were tested[1] with different feature sets to explore the space of possibilities and compare performances – table III. The parameter C of the classifier controls the tradeoff between misclassifying the data and the expected capacity of generalization.

The application of the ordinal model on the working data reveals a better generalization capability than the standard

[1]Experiments were carried out in Matlab, using the Support Vector Machine toolbox for Matlab, vs 2.51, by Anton Schwaighofer, to train the SVM standard model.

Feature set			multiclass SVM		ordinal SVM	
			# errors	# errors	# errors	# errors
Arms	Asy	Scar, colour	$C = 10$	$C = 1000$	$C = 10$	$C = 1000$
Down	BRA	Mean	(04.83; 0.19)	(03.81; 0.19)	(08.09; 0.16)	(08.77; 0.18)
		Median	(07.61; 0.19)	(07.53; 0.22)	(13.09; 0.24)	(11.67; 0.23)
	LBC	Mean	(05.59; 0.14)	(01.85; 0.15)	(05.98; 0.09)	(04.09; 0.11)
		Median	(04.88; 0.14)	(02.93; 0.14)	(06.00; 0.19)	(05.95; 0.18)
	UNR	Mean	(06.71; 0.21)	(05.61; 0.27)	(06.33; 0.18)	(08.35; 0.18)
		Median	(08.49; 0.19)	(06.58; 0.20)	(07.95; 0.16)	(07.98; 0.18)
	pBRA	Mean	(06.58; 0.18)	(03.85; 0.17)	(08.30; 0.19)	(08.07; 0.18)
		Median	(07.75; 0.23)	(07.69; 0.29)	(12.96; 0.23)	(13.68; 0.25)
	pLBC	Mean	(05.76; 0.14)	(04.61; 0.22)	(06.88; 0.11)	(06.86; 0.12)
		Median	(06.64; 0.16)	(04.78; 0.14)	(09.39; 0.23)	(09.35; 0.16)
	pUNR	Mean	(07.78; 0.21)	(07.47; 0.24)	(08.04; 0.21)	(07.96; 0.19)
		Median	(08.73; 0.25)	(06.85; 0.25)	(12.23; 0.23)	(13.26; 0.23)
Up	BRA	Mean	(11.32; 0.25)	(10.37; 0.22)	(09.65; 0.23)	(08.98; 0.19)
		Median	(13.05; 0.31)	(12.98; 0.34)	(14.00; 0.33)	(13.26; 0.30)
	LBC	Mean	(09.34; 0.19)	(08.32; 0.22)	(10.33; 0.18)	(09.84; 0.18)
		Median	(10.15; 0.24)	(09.42; 0.24)	(10.21; 0.16)	(10.56; 0.18)
	UNR	Mean	(07.54; 0.15)	(07.51; 0.17)	(07.77; 0.16)	(07.21; 0.16)
		Median	(13.05; 0.34)	(11.97; 0.38)	(13.91; 0.33)	(17.91; 0.39)
	pBRA	Mean	(10.36; 0.25)	(07.59; 0.25)	(09.44; 0.23)	(08.96; 0.21)
		Median	(12.56; 0.27)	(12.61; 0.27)	(19.11; 0.32)	(17.56; 0.32)
	pLBC	Mean	(08.53; 0.20)	(09.24; 0.22)	(08.63; 0.21)	(08.09; 0.21)
		Median	(08.53; 0.17)	(08.54; 0.24)	(11.04; 0.21)	(13.61; 0.25)
	pUNR	Mean	(06.76; 0.15)	(07.37; 0.20)	(06.91; 0.14)	(07.40; 0.14)
		Median	(13.22; 0.29)	(12.24; 0.32)	(15.14; 0.26)	(17.49; 0.32)

TABLE III

AVERAGE OF TRAIN AND TEST ERRORS.

multiclass SVM approach. Although the train error rate is higher than with standard SVM, the expected error rate over unseen patients attains the lowest values with this model.

From table III, we also conclude that the best results are attained with arms down. It is also apparent the superiority of LBC over the other asymmetry measures under study to discriminate classes. It is worth pointing that the feature set with LBC is fairly robust against using the mean or median of the scar visibility and colour change. This suggests that, when these two attributes are replaced by objective measures, the performance of the classifier will not be severely affected. Due to the small dataset available, the ordering of the feature sets should be taken with caution: a dataset of only 60 unique individuals may be insufficient to yield statistically significant results.

It is also instructive to identify which patients were incorrectly classified. This is illustrated graphically in figure 7 for the (*LBC*, *arms down* / *Scar* / *Colour*) feature set, with $C = 10$ (points highlighted with a star marker), and in a tabular form for a few more feature sets in table IV. It is interesting to note that most of the misclassified cases are located in the transition between two classes in the feature space and correspond to individuals with a low agreement value in the answers of the Delphi panel [1].

Finally, figure 8 shows the parallel planes (boundaries) separating each pair of classes, for the (*LBC*, *arms down* / *Scar* / *Colour*) feature set, with $C = 1000$, and table IV identifies which patients were incorrectly classified. The decision rule to classify unseen patients, using the boundary planes outputted by the ordinal classifier, can be cleanly stated as:

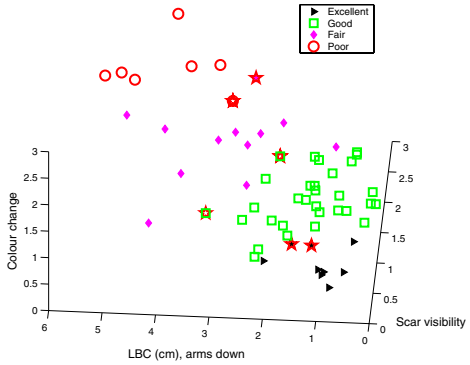


Fig. 7. misclassified points for LBC, arms down.

Feature set arms down	misclassified points	
	$C = 10$	$C = 1000$
LBC, mean	#1 #5 #6 #22 #36 #60	#5 #6 #22 #60
LBC, median	#5 #6 #22 #28 #49 #60	#5 #6 #22 #28 #49 #60

TABLE IV

MISCLASSIFIED PATIENTS, FOR THE ORDINAL MODEL.

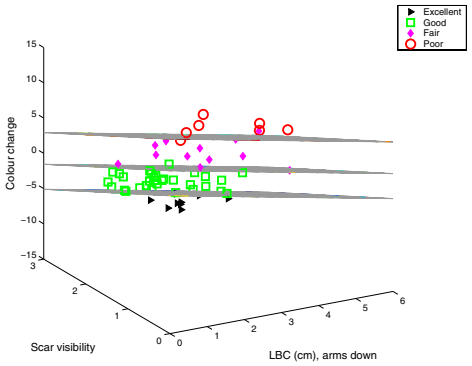


Fig. 8. Decision boundaries, for the ordinal model.

$$\text{if } \left(\; 0.33 \text{ LBC} + 1.00 \text{ Scar} + 0.31 \text{ Colour} \; \right)$$

$$\begin{cases} < 1.38 & \Rightarrow \text{ Excellent} \\ > 1.38 \wedge < 2.52 & \Rightarrow \text{ Good} \\ > 2.52 \wedge < 3.92 & \Rightarrow \text{ Fair} \\ > 3.92 & \Rightarrow \text{ Poor} \end{cases}$$

IV. Discussion

We have presented a comprehensive and principled methodology to classify semi-objectively the aesthetic result of the conservative breast cancer treatment. Our experiments have shown that it is possible to construct such an algorithm with an acceptable low error rate. The novel methodology described in this report draws from a diverse body of knowledge, borrowed mainly from the pattern classification community. Armed with these tools, we have corroborated early assumptions that the overall cosmetic result is mainly the contribution of the breast asymmetry, colour skin change and scar visibility.

From the conducted experiments, the main conclusions are that the position arms down clearly provide the best asymmetry features, with the lower breast contour outperforming both the breast asymmetry measure and the upward nipple retraction. The SVM ordinal model attained the best error rate and generalization capability.

Several unresolved issues are raised by the conducted research. The discarded patient, considered as an outlier, was the only non-caucasian patient in the set. This, together with the reduced set of patients available, raises doubts about the universality of the model. Further tests have to be conducted on a larger set of patients. This is the topic of ongoing research. Other directions in future work include the definition of objective measures for the skin colour change and surgical scar visibility. Software, to automate the evaluation process, is also currently being developed.

References

[1] M. J. Cardoso, A. C. Santos, J. Cardoso, H. Barros, and M. C. Oliveira, "Choosing observers for evaluation of aesthetic results in breast cancer conservative treatment," *International Journal of Radiation Oncology, Biology and Physics*, vol. 61, pp. 879–881, 2005.

[2] J. R. Harris, M. B. Levene, G. Svensson, and S. Hellman, "Analysis of cosmetic results following primary radiation therapy for stages i and ii carcinoma of the breast," *International Journal of Radiation Oncology Biology Physics*, vol. 5, pp. 257–261, 1979.

[3] R. D. Pezner, M. P. Patterson, L. R. Hill, N. Vora, K. R. Desai, J. O. Archambeau, and J. A. Lipsett, "Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 11, pp. 575–578, 1985.

[4] D. Clarke, A. Martinez, and R. S. Cox, "Analysis of cosmetic results and complications in patients with stage i and ii breast cancer treated by biopsy and irradiation," *International Journal of Radiation Oncology Biology Physics*, vol. 9, pp. 1807–1813, 1983.

[5] G. F. Beadle, S. Come, I. C. Henderson, B. Silver, S. Hellman, and J. R. Harris, "The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 10, pp. 2131–2137, 1984.

[6] R. D. Pezner, J. A. Lipsett, N. L. Vora, and K. R. Desai, "Limited usefulness of observer-based cosmesis scales employed to evaluate patients treated conservatively for breast cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 11, pp. 1117–1119, 1985.

[7] E. V. Limbergen, E. V. Schueren, and K. . V. Tongelen, "Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. proposal of a quantitative scoring system," *Radiotherapy and oncology*, vol. 16, pp. 159–167, 1989.

[8] L. I. Tsouskas and I. S. Fentiman, "Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer," *Breast Cancer Research and Treatment*, vol. 15, pp. 185–190, 1990.

[9] C. Vrieling, L. Collette, E. Bartelink, J. H. Borger, S. J. Brenninkmeyer, J. C. Horiot, M. Pierart, P. M. Poortmans, H. Struikmans, E. V. der Schueren, J. A. V. Dongen, E. V. Limbergen, and H. Bartelink, "Validation of the methods of cosmetic assessment after breast-conserving therapy in the eortc "boost versus no boost trial," *International Journal of Radiation Oncology Biology Physics*, vol. 45, pp. 667–676, 1999.

[10] M. Noguchi, Y. Saito, Y. Mizukami, A. Nonomura, N. Ohta, N. Koyasaki, T. Taniya, and I. Miyazaki, "Breast deformity, its correction, and assessment of breast conserving surgery," *Breast cancer research and treatment*, vol. 18, pp. 111–118, 1991.

[11] S. K. Al-Ghazal, L. Fallowfield, and R. W. Blamey, "Patient evaluation of cosmetic outcome after conserving surgery for treatment of primary breast cancer," *European journal of surgical oncology*, vol. 25, pp. 344–346, 1999.

[12] S. K. Al-Ghazal, R. W. Blamey, J. Stewart, and A. L. Morgan, "The cosmetic outcome in early breast cancer treated with breast conservation," *European journal of surgical oncology*, vol. 25, pp. 566–570, 1999.

[13] L. Krishnan, A. L. Stanton, C. A. Collins, V. E. Liston, and W. R. Jewell, "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*, vol. 91, pp. 2282–2287, 2001.

[14] J. Jones and D. Hunter, "Consensus methods for medical and health services research," *British Medical Journal*, vol. 311, pp. 376–380, 1995.

[15] F. Hasson, S. Keeney, and H. McKenna, "Research guidelines for the delphi survey technique," *Journal of Advanced Nursing*, vol. 32, pp. 1008–1015, 2000.

[16] M. J. Cardoso, J. Cardoso, A. C. Santos, H. Barros, and M. C. Oliveira, "Interobserver agreement and consensus over evaluation of breast cancer conservative treatment," *The Breast (accepted for publication)*, 2005.

[17] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley Interscience, 2001.

[18] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," in *Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2000, pp. 9–16.

[19] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., vol. 10. The MIT Press, 1998.

[20] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Neural Information and Processing Systems (NIPS)*, 2002.