Evaluation of border irregularity in pigmented skin lesions against a consensus of expert clinicians

Ela Claridge[a], Jon D Morris Smith[a] and Per N Hall[b]

[a]School of Computer Science, The University of Birmingham, Birmingham B15 2TT
[b]Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ

Abstract. The irregularity of skin lesion borders is a significant diagnostic factor when assessing a lesion for malignancy. A ranking experiment involving irregularity assessment showed poor average ranking correlation among 20 experienced clinicians (r=0.44). However, for 50% of the group correlation was high (r=0.81), indicating high degree of consensus. A quantitative measure of irregularity based on fractal dimension, derived by computer image analysis, corresponded well to the assessment of this sub-group (r=0.88) but not to the assessment of the entire group (r=0.3). It is proposed that only consensual data should be used as the basis for development and evaluation of medical image analysis algorithms.

1. Introduction

Images provide an important input to clinical diagnosis however, in many medical domains their interpretation is difficult. For example, in the diagnosis of pigmented skin lesions, which include potentially fatal malignant melanoma, it has been shown that reliable clinical diagnosis was only achieved when there was a consensus of three expert clinicians [1]. The difficulty is even greater for general practitioners who are normally the first point of referral for a patient anxious about a skin lesion; an average British GP will see only one malignant lesion in four years. The only way to diagnose a lesion with absolute certainty is to carry our a biopsy. To biopsy all suspected lesions is impractical and also unpleasant to a patient and expensive to NHS.

As a step towards improving diagnostic performance, a number of diagnostic checklists have been proposed, including UK "Seven Point Checklist" [2] and the American "ABCDE" list. The lists specify visual features associated with malignant lesions, including asymmetry, irregularity of border outline, irregularity of the colour and changes in lesion appearance and size. On first sight a list may seem to provide definite clues as to diagnosis. In reality, however, interpretation of these visual features is inconsistent both within and among observers [3].

Computer based image interpretation, offering objectivity and repeatability, seems to be an attractive proposition. Many systems developed to date (see [4] and [5] for reviews) apply standard measures and techniques to quantify the visual features and then subject the derived measurements to statistical analysis to establish whether any correlation exists between a measure and the diagnosis. The authors of this paper have been following a different approach, where a measure to characterise a given visual feature is selected on the basis of its agreement with the consensus of expert clinicians. Top experts *do* use visual features successfully for diagnosis. If an algorithm interprets the signs in the same way as a top expert, it is likely to be able to provide basis for an equally competent diagnosis. As with most tasks requiring quantification of visual features in natural images, the problem with this approach is that "the correct" interpretation is difficult or impossible to define in absolute terms. Unfortunately, finding a consensus is also a non-trivial task.

This paper is concerned with quantitative assessment of irregularity of boundary of skin lesions. On the practical level, the purpose of work described here was to establish whether a measure selected to describe irregularity is consistent with a consensus of expert clinicians. A further objective was to contribute to development of a general methodology for using image-related data provided by clinical experts to develop and evaluate computer image analysis algorithms.

2. Border irregularity

The irregularity of skin lesion borders was determined to be the most significant diagnostic factor when assessing a lesion for malignancy [6]. Furthermore, the analysis of the vocabulary used to describe lesion features revealed that, when describing malignant melanoma, significant emphasis is placed on border irregularity [7]. These results, coupled with the presence of border irregularity as a major feature in the Seven Point Checklist, indicate that this feature is an important factor in the diagnostic process. Early experiments by the authors [8] showed that clinicians frequently had difficulty in consistently assessing the border irregularity of lesion outlines; for example, the same lesion border, when rotated of flipped, was judged to have different degree of irregularity. An

irregularity measure based on the fractal dimension developed along that experimental work was consistent and repeatable and its sensitivity and specificity in classifying lesions into malignant and benign was 74% and 75% respectively.

The fractal dimension was chosen as a measure of irregularity because it was shown to correspond to the assessment of border irregularity by humans [9]. Prior to that, Tamura et al [10] demonstrated 98% correlation between shapes ranked by ascending fractal dimension and the ranking performed by naive subjects. The subsequent application of this method to pigmented skin lesions showed that probability of a lesion being malignant increased with increasing fractal dimension, and hence border irregularity [8]. The fractal dimension as a measure of border irregularity is by no means the only metric by which this feature can be assessed. Examples of other measures include the ratio of the circumference to the circumference of a circle with the same area [11]. A similar measure calculates the ratio of the area of the lesion to that of a circle with the same circumference [12]. These two, and similar methods, represent an approach to quantification of visual features which may provide a useful result but make use of methods without any indication of their relationship to the assessment of that feature by clinicians.

3. Methods and results

3.1 Ranking experiment

In order to establish whether the fractal dimension does indeed reflect the notion of "border irregularity" used by clinicians, an experiment was set up in which the ranking of lesion borders by the clinicians was compared to ranking according to magnitude of the fractal dimension. A set of twenty pigmented skin lesions with known clinical and histological diagnosis was selected by a clinical expert (PNH). The set comprised 7 malignant lesions (4 true-positives and 3 false-negatives) and 13 non-malignant ones (10 true-negatives and 3 false-positives) and thus contained lesions which are easy as well as those which are difficult to diagnose by clinical examination. The clinical expert has manually outlined the lesions and the outlines were stored as digital images and also printed on cards, one outline per card. Figure 1 shows examples of the outlines.

Twenty expert clinicians, drawn from the Melanoma Study Group and from specialist dermatology clinics in the West Midlands, took part in the experiment. The subjects were given the randomly shuffled pack of the cards containing the lesion outlines and were asked to sort the cards in the order of ascending irregularity of the border outline. The results were recorded by noting the sequence of numbers shown on the cards. The fractal dimension (FD) of the outlines was computed using a method based on the algorithm by Flook [13], described in [8]. The subsequent ranking used the "structural" component of FD which corresponded to the ruler length from 15 to 42 pixels (1.4 to 3.8 mm). The fractal dimension of the lesions in the set varied from 1.002 to 1.159.

1.027 1.068 1.088 1.151

Figure 1. Examples of lesion outlines together with their "structural" fractal dimension.

3.2 Results

The results of the experiment involving clinicians were analysed to assess the degree of consensus among the clinicians and also to assess the degree of correspondence between the ranking produced by the clinicians and the fractal dimension. The degree of consensus was measured using the average rank correlation (r_s) derived from the Kenall tau coefficient of concordance [14]. Given the number of subjects (m) and the number of samples to be ranked (N), the tau coefficient (W) is calculated thus:

$$W = \frac{12\Sigma_j T_j^2}{m^2 N(N^2 - 1)} - \frac{3(N + 1)}{N - 1}$$

where T_j is the sum total of ranks for sample j.
The average rank correlation is computed as:

$$r_s = \frac{mW - 1}{m - 1}$$

For 20 subjects and 20 samples, the average rank correlation r_s was 0.44 (W = 0.47), suggesting relatively poor correlation among the clinicians (perfect agreement would produce r_s=1.0). These results confirmed earlier findings that clinicians are inconsistent in their assessment of irregularity [6], [8].

To compare the degree of correspondence between the ranking produced by the clinicians and the fractal dimension, two sets of statistical tests were performed. In the first one, the ranking according to value of the fractal dimension was compared with the average ranking given by the clinicians to each outline. In this instance, the average rank correlation was r_s = 0.3 (W = 0.65), indicating poor correlation. A similar experiment using median instead of average gave a nearly identical result. At a first glance these results were disappointing, suggesting that, contrary to psychophysical evidence, the association between perceived irregularity of a boundary and its fractal dimension did not show in this particular domain. However, the low correlation could simply reflect the poor agreement among the clinicians.

To test this hypothesis, the second test was carried out where the ranking associated with the fractal dimension was compared with the rankings of individual clinicians. The value of r_s was found to vary between 0.27 and 1.0 (W from 0.63 to 1.0, respectively), with 50% results having r_s > 0.5. This meant that the rankings of approximately half of the subjects were in reasonable agreement with the fractal dimension. The question was, whether those subjects were also in agreement with each other.

In a subsequent test, the average rank correlation values were calculated for each pair of the subjects (210 pairs) and the average score was calculated for each subject to estimate the degree of consensus between that subject and all the others. The subjects were then sorted in order of increasing consensus (Figure 2). Interestingly, the top 50% subjects were the same as those whose ordering agreed most with ordering by the fractal dimension. The average rank correlation among the top 50% was 0.81 (W = 0.83); for the other sub-group it was 0.21 (W = 0.29). Thus consistency among the top half of the subjects was much greater than among the bottom half.

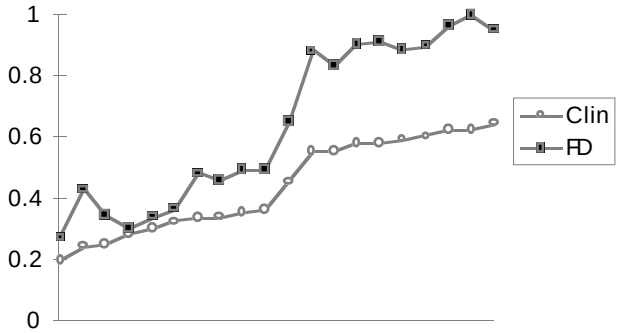

Figure 2: The plot shows for each clinician: the average rank correlation score showing agreement in ranking between the given clinician and all the others (open circles); and the rank correlation value showing agreement in ranking between the given clinician and ranking by fractal dimension (filled squares).

When the correlation between the average rank number and the fractal dimension was computed for these two groups, r_s was 0.88 for the top group (W = 0.94) and 0.39 for the other group (W = 0.69). The results for the top group support an earlier research [10] and indicate that the fractal dimension is a good candidate for a measure of irregularity which is consistent with the human assessment *when the consensus exists for the human data*.

4. Discussion and further work

On the basis of the results obtained above, it is possible to conclude that the fractal dimension as a measure of border irregularity agrees well with the assessment of a sub-group of experimental subjects. More interestingly, however, there is also a strong ranking correlation *among* the subjects within this sub-group, indicating that their assessment is mutually consistent. By contrast, the rankings of the other sub-group do not show consistency.

These observations suggests a possible methodology for evaluating algorithms which aim to quantitatively characterise visual features in a manner consistent with assessment by a consensus of observers. After collecting experimental data, its (maximal) mutually consistent sub-set should be found. The degree of consistency required could be adjusted by changing a suitable consistency threshold (e.g., r_s in the experiments above) and by specifying the minimum (percentage) size of the subset with a given threshold. The results of an algorithm can be then compared against the results of the internally consistent subset.

If the comparisons are made against the average result of all the subjects, an apparent lack of agreement does not necessarily indicate that the algorithm is inappropriate *if the consensus among the subjects is poor*. Poor consensus can arise for many reasons, for example through misunderstanding of a task by a number of subjects, through lack of concentration or lack of application, because a subset of images is particularly problematic, or because a unique answer does not exist. Whatever the reason, using such inconsistent data as a ground truth for developing and training of computer algorithms is not likely to result in robust and acceptable solutions.

This work forms a part of research at the University of Birmingham into developing theories and techniques which aim to aid clinicians in early diagnosis of malignant melanoma. Further work will proceed along several lines. The methodology for finding consensual subset in data generated by clinical colleagues will be formalised and extended to include types of assessment other than ranking and applied to other diagnostic features such as pigment variation, symmetry, notchiness and edge definition. The computer-derived measures which are shown to be consistent with expert assessment will be applied to a large database of lesion images to determine their correlation with diagnosis. The images with associated computer measures characterising diagnostically important visual features will be used for training and decision support using a framework of the MEDIATE system which has been shown to be successful in training radiologists to in interpret MRI images [15].

References

1. R. Curley, M. Cook & Fallowfield. “Accuracy in clinically evaluating pigmented lesions”, *British Medical Journal* **299**, pp 16-18, 1989.
2. R. MacKie. *An illustrated guide to the recognition of early malignant melanoma*. University Department of Dermatology, Glasgow, 1985.
3. E. Higgins, P.N. Hall, P. Todd, R. Murthi & A. du Vivier. “The application of the seven-point check-list in the assessment of benign pigmented lesions”, *Clinical and Experimental Dermatology* **17**, 313-315, 1992.
4. W.V. Stoecker & R.H. Moss. “Digital imaging in dermatology”, *Computerized Medical Imaging and Graphics* **16**(3), 145-150, 1992.
5. P.N. Hall, E. Claridge & J.D. Morris Smith. “Computer screening for early detection of melanoma - is there a future?” *British Journal of Dermatology* **132**, 325-338, 1995.
6. M. Keefe, D. Dick & R. Wakeel. “A study of the value of the seven-point checklist in distinguishing benign pigmented lesions from melanoma”, *Clinical and Experimental Dermatology* **15**, 167-171, 1990.
7. J.D. Morris Smith *Characterisation of the appearance of pigmented skin lesions by computer methods compatible with clinical assessment*. PhD Thesis, The University of Birmingham, School of Computer Science, 1997.
8. E. Claridge, P.N. Hall, M. Keefe & J. Allen. “Shape analysis for classification of malignant melanoma”, *Journal of Biomedical Engineering* **14**, 229-234, 1992.
9. A. Pentland. “Fractals: A model for both texture and shading”. In Hoffman D (Ed) *Annual Meeting of the Optical Society of America*: MIT, 1984.
10. H. Tamura, S. Mori & T. Yamawaki. “Textural features corresponding to visual perception”, *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-8**, 460-473, 1978.
11. R. White, D. Rigel & R. Friedman. “Computer applications in diagnosis of malignant melanoma”, *Melanoma Skin Cancer Update* **9**(4), 695-702, 1991.
12. J. Golston, W.V. Stoecker, R. Moss & I. Dhillon. “Automatic detection of irregular borders in melanoma and other skin tumours”. *Computerized Medical Imaging and Graphics* **16**, 199-203, 1992.
13. A.G. Flook. “The use of dilation logic on the Quantimet to achieve fractal dimension characterization of texture and structured profiles”. *Powder Technology* **21**, 295-298, 1978.
14. W. Hays. *Statistics*. Chapter 19: Holt, Rinehart and Winston, 1988.
15. M. Sharples, N. Jeffery, D. Teather *et al.* “A socio-cognitive engineering approach to the development of a knowledge-based training system for neuroradiology”. In du Boulay B & Mizoguchi R (Eds) *AI in Education*, 402-409: AIO Press, 1997.