Skin Research and Technology 2011; 17: 35–44
Printed in Singapore · All rights reserved
doi: 10.1111/j.1600-0846.2010.00460.x

Skin Research and Technology

Weighted performance index for objective evaluation of border detection methods in dermoscopy images

Rahil Garnavi[1], Mohammad Aldeen[1] and M. E. Celebi[2]

[1]NICTA Victoria Research Laboratory, Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Vic., Australia and [2]Department of Computer Science, Louisiana State University, Shreveport, LA, USA

Purpose: This paper presents a novel approach for objective evaluation of border detection in dermoscopy images of melanoma.
Background: In melanoma studies, border detection is a fundamental step toward the development of a computer-aided diagnosis system. Therefore, its accuracy is essential for accurate implementation of the subsequent parts of the diagnostic system.
Method: An objective evaluation procedure of border detection methods is presented. The evaluation procedure uses the weighted performance index, which is composed of weighted metrics of sensitivity, specificity, accuracy, precision, border error and similarity. This index can also be used to optimize the parameters of a border detection method.
Result and conclusion: Experiments are performed on 55 high-resolution dermoscopy images. Using the union of four sets of dermatologist-drawn borders as the ground truth, weighted metrics of sensitivity, specificity, accuracy, precision, border error and similarity are evaluated. Then, the weighted performance index is constructed and used to optimize the parameters of the hybrid border detection method. The outcome of the optimization process, verified through statistical analysis, yields a higher degree of agreement between automatic borders and the ground truth, compared with using standard metrics only. Finally, the weighted performance index is used to evaluate five recently reported border detection methods.

Key words: computer-aided diagnosis – melanoma – dermoscopy – border detection – weighted performance index – evaluation metric

Accepted for publication 21 June 2010

Malignant melanoma is one of the most lethal and rapidly increasing cancers. It represents 10% of all cancers in Australia, and its per-capita incidence is four times higher than that in Canada, the United Kingdom and the United States, with more than 10,000 cases diagnosed and around 1700 deaths annually (1).

Dermoscopy (also known as epiluminescence microscopy) is a non-invasive *in vivo* imaging technique, which allows for a magnified and clear visualization of the morphological structures of the skin that are not visible to the naked eye. With the use of dermoscopy and dermoscopic algorithms (2, 3), such as pattern analysis, ABCD rule of dermoscopy, Menzies method, seven-point checklist, and the CASH algorithm, which have benefited from the technological advancements over the past few decades, the diagnosis of melanoma has been improved compared with the simple naked-eye examination between 5% and 30% depending on the type of skin lesion and the experience of the dermatologist (4).

However, clinical diagnosis of melanoma is inherently subjective and its accuracy has been an issue of concern (4). With the goal of removing subjectivity and uncertainty from the diagnostic process and providing a reliable second independent opinion to dermatologists, computer-based analysis of dermoscopy images has become a major research area.

Border detection is a fundamental step toward the development of a computer-aided diagnosis of melanoma, which involves separating the lesion from the background skin. The accuracy of the detected border is essential for accurate implementation of the subsequent parts of the diagnostic system, i.e. feature extraction and classification, because of the fact that features such as asymmetry and border irregularity are highly dependent on border detection. In addition, the lesion inside the detected border reveals information about homogeneity, dermoscopic patterns and lesion color.

Detecting the lesion borders in dermoscopy images is considered to be challenging (5) due to (1) the low contrast between the lesion and the background skin, (2) presence of unwanted artifacts within the image such as hairs, skin lines, blood vessels, air bubbles, black frames, camera scale and blue/purple surgical markings, (3) fuzziness of the border and (4) the color variegation within the lesion.

Numerous border detection methods have been reported in the literature (5). Recent methods include histogram thresholding (6), thresholding followed by region growing (7), color clustering (8, 9), statistical region merging (10), JSEG algorithm based on color quantization and spatial segmentation (11), two-stage k-means ++ clustering followed by region merging (12), global thresholding on optimized color channels followed by morphological operations (13) and hybrid thresholding (14).

On the other hand, objective evaluation of border detection methods has not been explored in depth (5). Existing evaluation methods are either through visual assessment of the detected borders by dermatologists, which suffers from subjectivity, or through an objective evaluation, where the closeness of an automatic border produced by the border detection method is compared with that manually drawn by dermatologists. With respect to objective evaluation methods, different metrics have been used, namely, sensitivity, specificity, accuracy, border error, similarity, precision, pixel misclassification probability (15) and normalized probabilistic rand index (16).

There exist two problems in the evaluation of the border detection methods: First, the abovementioned metrics are generic and have been widely applied in different domains. However, in the application to border detection of dermoscopy images of melanoma lesions, it is crucial that dermatologists' perspectives are taken into account, which raises the need to customize the standard metrics to reflect their respective practical importance in the evaluation process. Second, it is often the case that a border detection method yields a superior result according to one evaluation metric, yet is defeated by other methods with respect to another metric(s). In other words, there is no comprehensive metric for comparing different automated methods. In this paper, we propose a novel approach to tackle these two problems.

The rest of the paper is organized as follows. The standard metrics are reviewed in 'Standard Evaluation Metrics'. The proposed evaluation approach is presented in 'Proposed Evaluation Metrics'. The experimental results are discussed in the penultimate section. The last section provides the conclusion.

Standard Evaluation Metrics

The standard metrics of sensitivity, specificity, accuracy, border error, similarity and precision [expressed in Eqs (1)–(6)] are statistical measures based on true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Sensitivity shows the percentage of the actual lesion that has been detected accurately by the automated method. Specificity shows the percentage of the actual background skin that has been detected accurately by the automated method. Precision shows what percentage of the detected border is the true lesion. Accuracy and similarity are two other metrics that exhibit the degree of agreement between the automatic border produced by an automated method and the manual border drawn by dermatologists (the gold standard), and border error measures the discrepancy between the two borders:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\% \tag{3}$$

$$\text{Similarity} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \times 100\% \tag{4}$$

$$\text{Border error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN}} \times 100\% \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{6}$$

Proposed Evaluation Metrics

The proposed evaluation approach aims to solve two problems: defining a comprehensive metric, which takes into account various evaluation metrics, and providing objective evaluation metrics that are meaningful in the context of a melanoma application. In the following, these two aspects are discussed.

Performance index (PI)

The performance of existing border detection methods has been commonly evaluated using either one evaluation metric (e.g. border error) or two metrics (e.g. sensitivity and specificity). As each of the metrics has its specific meanings and implications, a problem arises when it comes to objectively interpreting the results. This problem is compounded, especially when different metrics yield different rankings. Thus, the question of which method yields the best possible result when all of the evaluation metrics are considered simultaneously has not been resolved as yet. To determine this, we propose a comprehensive measure called PI for systematic evaluation of the performance of existing border detection algorithms. PI takes into account all of the standard metrics previously defined [Eqs (1)–(6)]:

$$\mathrm{PI} = \frac{\mathrm{Sn} + \mathrm{Sp} + \mathrm{Ac} + \mathrm{Sm} + \mathrm{Be} + \mathrm{Pr}}{6} \tag{7}$$

where Sn, Sp, Ac, Sm, Be and Pr refer to sensitivity, specificity, accuracy, similarity, 100% – border error and precision, respectively.

Weighted evaluation metrics

Objective evaluation of existing border detection methods requires metrics that are meaningful in the context of a melanoma application. Our experience with dermatologists has consistently shown that TP has the highest importance compared with the other three factors of TN, FP and FN, as dermatologists tend to have the entire lesion included in the automatic border. On the other hand, FP (the areas included by the automated method, yet excluded by the gold standard) has a minor degree of importance compared with FN (the areas excluded by the automated method, yet included by the gold standard). Accordingly, after exhaustive consultation with dermatologists, we attach a weighting of 1.5 to TP to indicate its overall importance. Furthermore, to emphasis the importance of FN over FP, in those metrics that include both FN and FP, we assign a factor of 0.5 to FP. As a result, the new set of equations is given by

$$W_{\mathrm{Sensitivity}} = \frac{1.5\mathrm{TP}}{1.5\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{8}$$

$$W_{\mathrm{Accuracy}} = \frac{1.5\mathrm{TP} + \mathrm{TN}}{1.5\mathrm{TP} + 0.5\mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \times 100\% \tag{9}$$

$$W_{\mathrm{Similarity}} = \frac{3\mathrm{TP}}{3\mathrm{TP} + \mathrm{FN} + 0.5\mathrm{FP}} \times 100\% \tag{10}$$

$$W_{\mathrm{Bordererror}} = \frac{0.5\mathrm{FP} + \mathrm{FN}}{1.5\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{11}$$

$$W_{\mathrm{Precision}} = \frac{1.5\mathrm{TP}}{1.5\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{12}$$

where $W_{\mathrm{Sensitivity}}$, for instance, stands for weighted sensitivity. The specificity metric does not contain the TP factor. Also, FP and FN do not appear simultaneously in this metric. Accordingly, it remains unchanged as in Eq. (2).

Weighted performance index (WPI)

Following the discussion presented in 'Performance index (PI)' a comprehensive metric is defined to objectively evaluate border detection methods for dermoscopy Images. It is called WPI and it takes into account the weighted metrics defined in Eqs (2) and (8)–(12). WPI is in fact a weighted average, where the weights are imbedded within each metric:

$$\mathrm{WPI} = \frac{W_{\mathrm{Sn}} + \mathrm{Sp} + W_{\mathrm{Ac}} + W_{\mathrm{Sm}} + W_{\mathrm{Be}} + W_{\mathrm{Pr}}}{6} \tag{13}$$

where W_{Sn}, Sp, W_{Ac}, W_{Sm}, W_{Be} and W_{Pr} refer to $W_{\mathrm{Sensitivity}}$, specificity, W_{Accuracy}, $W_{\mathrm{Similarity}}$, $100\% - W_{\mathrm{Border\ error}}$ and $W_{\mathrm{Precision}}$, respectively.

Experimental Results

The proposed evaluation method is tested on a set of 55 high-resolution dermoscopy images obtained from Royal Melbourne Hospital, Australia. The images were taken by professional photographers using a Canon EOS 450D camera under unified zooming and lighting conditions. They are 24-bit RGB color images with dimensions of 2000 × 1334 pixels in TIFF format.

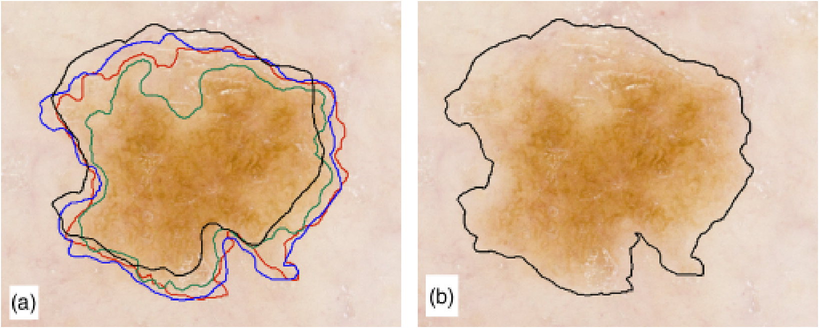

Fig. 1. *Gold standard for a sample dermoscopy image: (a) four different dermatologists-drawn borders and (b) the obtained union border.*

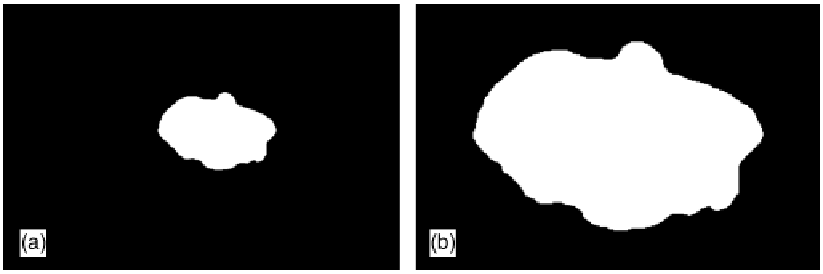

Fig. 2. *Segmentation result of a dermoscopy image (a) before and (b) after standardization.*

To validate the borders produced by existing methods, manual borders are independently drawn by four expert dermatologists using Wacom Intuos A4 size Tablet PC. We are aware that different approaches can be used to acquire the final ground truth, such as voting, averaging, etc. However, after discussion with dermatologists and considering the practical nature of the melanoma diagnosis, which calls for extreme caution when excluding portions of the image, and taking into account the inter-observer and intra-observer variability in borders drawn by dermatologists, we calculate the union of the four manually drawn borders for each image and consider that as the final ground truth. This choice is in line with the emphasis we place on the TP factor in the proposed evaluation metrics. Figure 1 shows the four different dermatologist-drawn borders and the obtained union border for a sample dermoscopy image. As shown in Fig. 1, borders are close enough to each other, therefore, there would be no adverse effect for the way the ground truth is calculated

Standardization of the images

The lesion inside a dermoscopy image generally appears in different sizes and locations. The two metrics of accuracy and specificity include the TN factor, which refers to the number of pixels of the background skin that are accurately detected by the automated method. However, the TN factor depends considerably on the size of the lesion and its ratio to the whole image; thus, the value of the accuracy and specificity metrics are biased against images with small lesions. To the best of our knowledge, this issue has not been addressed in previously published studies.

To balance the effect of large TN and normalize the accuracy and specificity metrics, we set a frame of background skin around the lesion, such that the area of the rectangular image frame is twice as large as that of the smallest imaginary rectangle enclosing the lesion, with horizontal and vertical sides. This has the effect that the number of background pixels and lesion pixels is roughly the same. Figure 2 shows the segmentation result of a dermoscopy image before and after standardization.

Optimization of the parameters

As an advantage, the proposed WPI can be used to optimize automated border detection methods by tuning their parameters. For instance, in the hybrid thresholding method (14), two main parameters of window size (W) and bandwidth factor (B) are involved. W is the size of the window over which the local threshold is calculated, and the B shows the extent to which the initial border is expanded toward the background skin. For completeness, a summary of the method is provided in the following. Further details may be found in (14).

Hybrid thresholding method

Manual borders drawn by dermatologists tend to surround the borders produced by automated method (5, 14). As shown in Fig. 3, three areas are generally identified in dermoscopy images: core lesion, edge lesion and background skin. The width of the edge-lesion area can be quite variable depending on the skin color, lesion color and fuzziness of the border.

Forming the core lesion. As reported in (14), the core-lesion area is detected by applying global histogram thresholding to the optimal color channel of XoYoR obtained in the color optimization procedure (XoYoR combines X and Y color channels from the XYZ color space with R color channel from the RGB color space). This step includes the pre-processing operations of hair removal (17), noise filtering using a Gaussian low-pass filter (with a 10 × 10 kernel) and intensity adjustment. This is followed by application of the Otsu thresholding method (18) to the XoYoR color channel, and performing connected component analysis and morphological operations to obtain the initial border for the lesion and form the core-lesion area.

Forming the edge lesion. To expand or shrink the core-lesion boundary to the edge-lesion boundary, an adaptive local thresholding technique based on the Otsu method is used, where the histogram thresholding is applied to the X color channel determined as optimal in the color optimization procedure (13). Starting from an arbitrary point on the core-lesion boundary, the local threshold is calculated over a window of size *W*. If the local threshold value is less than a defined threshold called T_{expand}, the boundary is expanded by one pixel. If it is larger than a defined threshold called T_{shrink}, the boundary is shrunk. Otherwise, it is the *No Change* state, where based on the previous moves, a decision is made to either laterally move to the adjacent pixel or expand or shrink the boundary in a radial manner, along the line connecting the centroid of the lesion to the pixel on the core-lesion boundary. The lesion centroid is given by

$$(x_c, y_c) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right) \tag{14}$$

where *n* is the number of pixels along the border, and (x_i, y_i) is the coordinates of the *i*th lesion pixel. To define the threshold values for shrinkage and expansion [Eq. (16)], a bandwidth [Eq. (15)] is calculated based on background skin and core-lesion pixels values. The Otsu method is applied to the background skin area and core-lesion area to obtain estimates of these values.

$$\text{Bandwidth} = \%B \times (T_{\text{skin}} - T_{\text{coreLesion}}) \tag{15}$$

where *B*, the bandwidth factor, is the extent to which the core-lesion area is expanded toward the background skin. The local threshold is calculated for every pixel over its surrounding window and the process is stopped when the initial pixel is revisited.

$$T_{\text{expand}} = T_{\text{skin}} - \text{Bandwidth},\ T_{\text{shrink}} = T_{\text{skin}} + \text{Bandwidth} \tag{16}$$

To determine the optimal settings (*W*, *B*) for the method, we perform a comprehensive set of experiments on the image set of 55 high-resolution dermoscopy images, with *W* varying from 20 to 70 and *B* ranging from 10% to 90% (steps of 10). Consequently, 54 borders are obtained for each dermoscopy image. To evaluate the results, each border is compared with the ground truth and the standard and weighted metrics of sensitivity, specificity, accuracy, similarity, border error and precision, and the corresponding PI and WPI are calculated. In the following, the optimal pairs of *B* and *W* are obtained.

PI and WPI

PI and WPI are calculated for various *W* and *B* values over the image set using the value of the

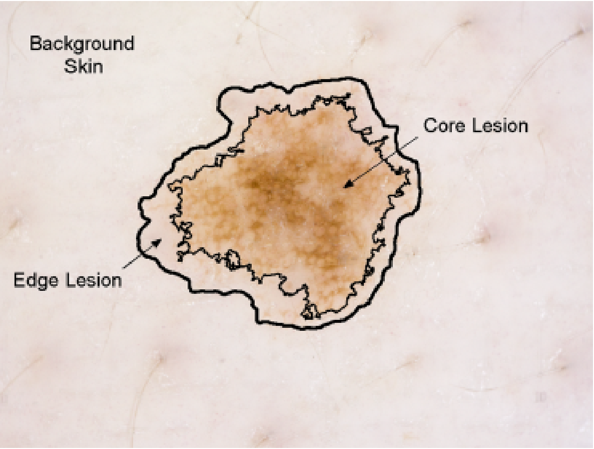

Fig. 3. Three different areas generally appear in dermoscopy images: core lesion, edge lesion and background skin.

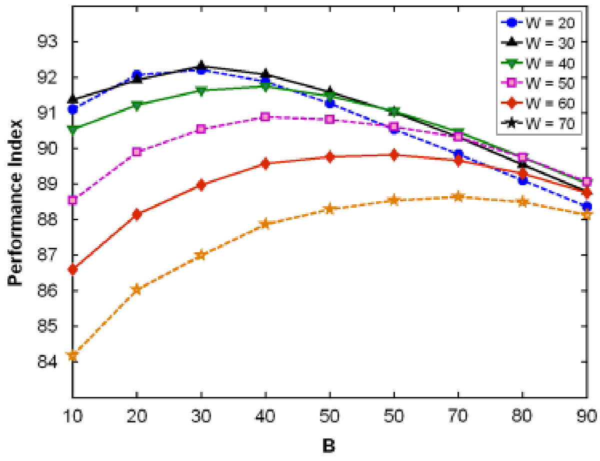

Fig. 4. Performance Index for various W *and* B *values over the image set.*

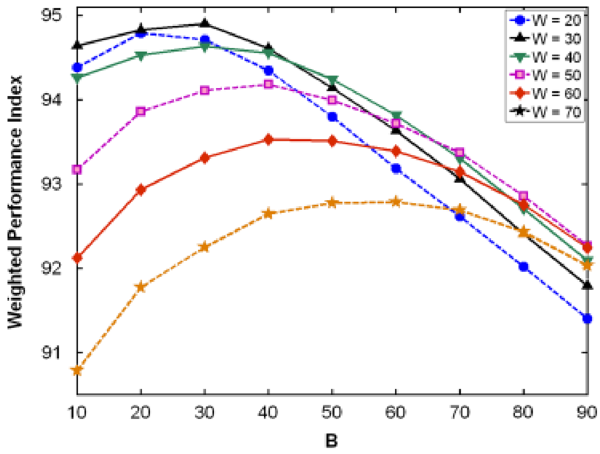

Fig. 5. Weighted Performance Index for various W *and* B *values over the image set.*

standard and weighted metrics of sensitivity, specificity, accuracy, similarity, border error and precision according to Eqs (7)–(13) and are averaged over 55 images. For each *W*, the family of mean PIs and mean WPIs vs. *B* is plotted, as shown in Figs 4 and 5. The two graphs of PI and WPI metrics are coherent and both reveal the optimal setting of (30, 30) for *B* and *W*, as shown in Figs 4 and 5. However, there is a distinction between the two analyses, i.e. standard vs. weighted metrics. The PI, which is based on standard metrics, yields a lower result than the WPI, which is based on proposed weighted metrics. For example, for W30B30, the mean value of PI is 92.30, whereas the mean value of WPI is 94.90. Thus, the weighted metrics, which are defined to reflect the dermatologists' perspectives, show a higher degree of agreement between automatic and manual borders, compared with standard metrics.

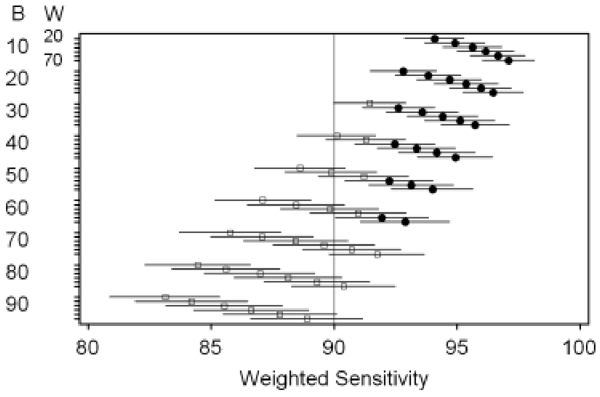

Fig. 6. Mean and 95% confidence interval of weighted sensitivity metric for various W *and* B *values over the image set.*

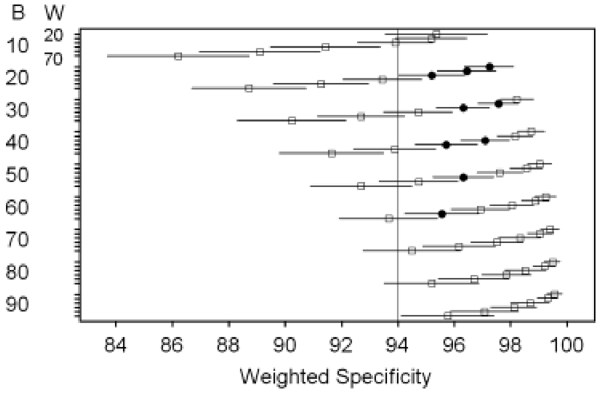

Fig. 7. Mean and 95% confidence interval of specificity metric for various W *and* B *values over the image set.*

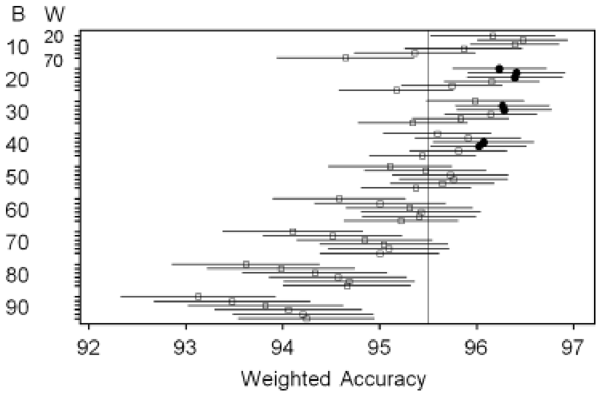

Fig. 8. Mean and 95% confidence interval of weighted accuracy metric for various W *and* B *values over the image set.*

Statistical analysis

Figures 6–11 show the mean value and 95% confidence interval (CI) for metrics of weighted sensitivity, specificity, weighted accuracy, weighted similarity, weighted border error and the weighted precision for various values of *W* and *B* parameters, respectively. We set levels of

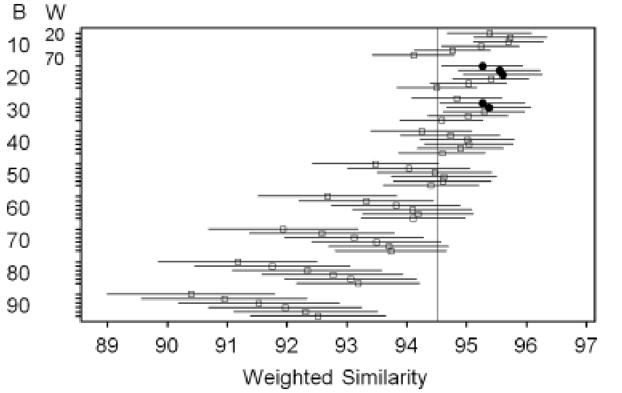

Fig. 9. Mean and 95% confidence interval of weighted similarity metric for various W *and* B *values over the image set.*

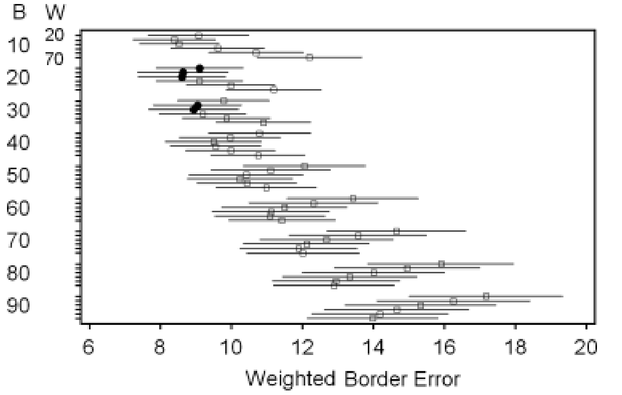

Fig. 10. Mean and 95% confidence interval of weighted border error metric for various W *and* B *values over the image set.*

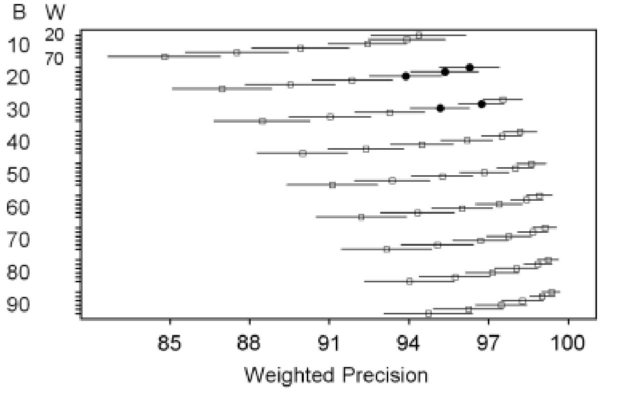

Fig. 11. Mean and 95% confidence interval of weighted precision metric for various W *and* B *values over the image set.*

acceptability for the lower bound of the CI. The levels are arbitrary but reasonable, and are a helpful guide for identifying acceptable parameter values across a range of metrics.

Owing to the importance of sensitivity, we start the analysis from this metric. As shown in Fig. 6, given the level of acceptability of 90% for weighted sensitivity, 26 sets of W and B, out of 54, are selected, which are marked by filled circles in the graph. Having the level of acceptability of 94% for specificity, nine sets from the previous 26 sets are nominated, as illustrated in Fig. 7. For the level of acceptability of 95.5% for weighted accuracy, as shown in Fig. 8, seven pairs of W and B meet the criteria.

As illustrated in Fig. 9, with the level of acceptability of 94.5% for weighted similarity, these seven pairs are further narrowed down to five sets of (20, 20), (30, 20), (40, 20), (30, 30) and (40, 30). Having these five sets, the border error metric is studied and as shown in Fig. 10, the above-mentioned five pairs are highly competitive where their respective mean and 95% confidence interval are very similar. To determine the optimal value for B and W, weighted precision metric is investigated. As shown in Fig. 11, the pair of (30, 30) achieves the best result, followed by (20, 20). This result is in accordance with the results obtained from the PI and WPI in the previous section.

Cross validation

In order to provide a stronger proof for the optimal B and W identified through the statistical analysis and the proposed WPI, we also perform a 11-fold cross validation process, where the data set of 55 images is iteratively partitioned into a 50-image subset for the train set and a five-image subset for test set, yielding 11 sets with unique combinations of test and train data. For each of the test sets, a family of WPI curves for different W and B is plotted, as illustrated in Fig. 12, which shows that all train sets converge to the value of (30, 30) for B and W, except for set 6. According to set 6, the optimal WPI is ($W = 30$, $B = 20$), yet its resulting WPI is almost similar to the WPI value for ($W = 30$, $B = 30$). Table 1 shows the resultant WPI for images and the corresponding mean and standard deviation for each test set using the optimal setting of ($W = 30$, $B = 30$), which demonstrates the acceptability of the identified setting.

Comparison between automated methods

As another advantage, the WPI has also been used to evaluate five recent border detection methods applied to the image set of 55 dermoscopy images (see Fig. 13). These methods are dermatologist-like tumor extraction algorithm (DTEA) (7), JSEG (11), KPP (12), global thresholding on optimized color channel of XoYoR (13),

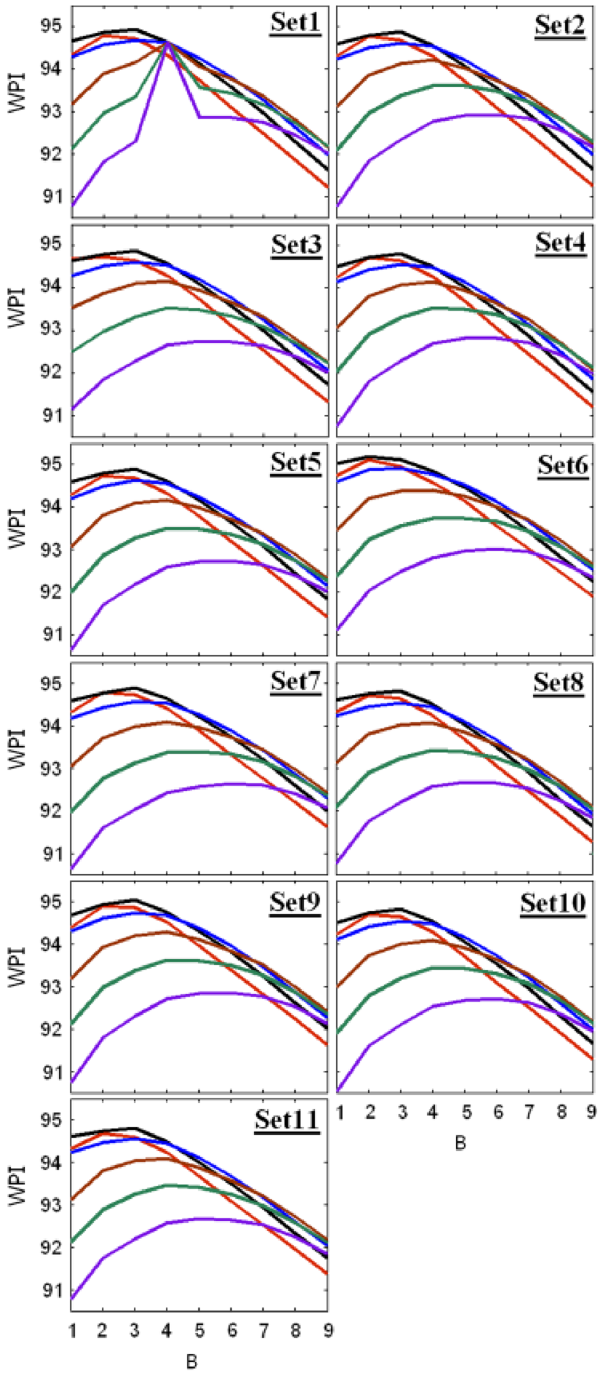

Fig. 12. Eleven-fold cross validation using weighted performance index evaluation metric.

and hybrid thresholding (14) with optimal parameters 'Hybrid (W30B30).'

Table 2 presents the 95% CI for the mean of the proposed metrics. As shown, a method may perform better than others with respect to some of the metrics, yet yields lower results with respect to others, e.g. the DTEA method achieves the highest specificity and weighted precision, yet it is overcome by the hybrid method with respect to weighted metrics of sensitivity, accuracy, similarity and border error. For these reasons, it has not been easy to provide an overall objective judgment as to which of the five methods is more suited for border detection of dermoscopy images. However, the use of the proposed WPI, as shown in the bottom row of Table 2, facilitates such a judgment, which can be made easily by comparing values of the calculated WPIs. According to the obtained

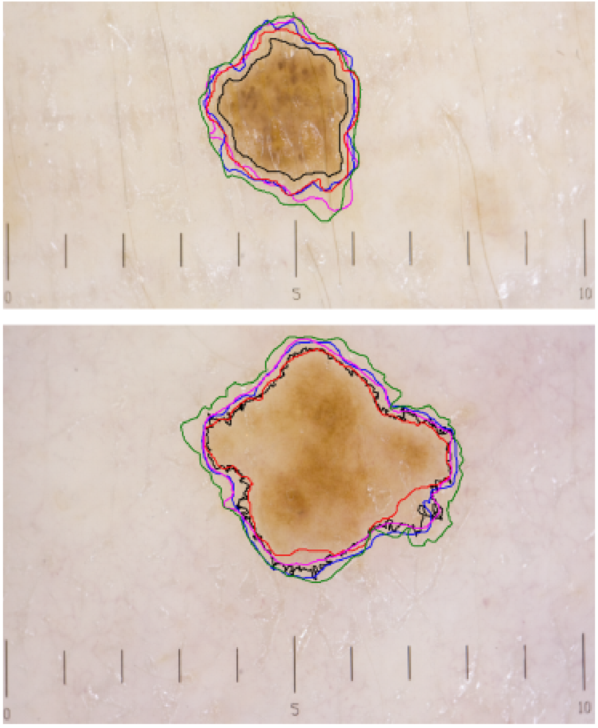

Fig. 13. Different automatic borders for two sample dermoscopy images.

TABLE 1. Weighted performance index for different images in each test set

Test	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
Image 1	94.29	94.78	97.13	96.95	93.50	95.55	95.34	95.67	93.91	95.91	96.95
Image 2	93.49	96.09	96.59	95.49	94.79	94.52	94.30	97.20	97.87	93.61	96.70
Image 3	95.22	94.80	95.29	94.78	95.10	87.83	96.42	95.32	89.95	96.62	91.97
Image 4	94.64	95.36	92.43	97.04	93.56	91.05	90.75	96.50	96.36	94.52	96.45
Image 5	94.80	94.78	94.51	94.58	97.66	95.10	98.02	93.49	89.51	97.15	97.39
Mean	94.49	95.16	95.19	95.77	94.92	92.81	94.96	95.64	93.52	95.56	95.89
Standard deviation	0.64	0.57	1.85	1.17	1.69	3.30	2.73	1.40	3.74	1.47	2.22

TABLE 2. Evaluation metrics (mean ± margin of error) for five border detection methods

	Hybrid (W30B30)	JSEG	DTEA	Global (XoYoR)	KPP
Weighted accuracy	96.27 ± 0.4	93.73 ± 0.9	92.49 ± 1.0	92.29 ± 0.7	91.28 ± 1.3
Weighted sensitivity	92.62 ± 1.4	84.48 ± 2.5	81.16 ± 2.6	80.89 ± 2.1	78.13 ± 3.7
Specificity	97.56 ± 0.7	99.59 ± 0.2	99.79 ± 0.1	99.63 ± 0.2	99.17 ± 0.6
Weighted precision	96.72 ± 0.8	99.43 ± 0.3	99.70 ± 0.1	99.45 ± 0.2	98.99 ± 0.7
Weighted similarity	95.26 ± 0.7	91.14 ± 1.6	89.20 ± 1.6	89.08 ± 1.4	86.76 ± 2.4
Weighted border error	9.04 ± 1.2	15.79 ± 2.5	18.98 ± 2.5	19.35 ± 2.1	22.39 ± 3.5
Weighted performance index	94.90 ± 0.5	92.09 ± 1.2	90.56 ± 1.2	90.33 ± 1.0	88.65 ± 1.7

TABLE 3. Mean ± margin of error of weighted performance index for comparisons of automated methods

	Hybrid (W30B30)	JSEG	DTEA	Global (XoYoR)	KPP
Hybrid (W30B30)	0	2.80 ± 1.0	4.34 ± 1.0	4.57 ± 0.8	6.24 ± 1.5
JSEG	− 2.80 ± 1.0	0	1.53 ± 1.1	1.76 ± 1.1	3.43 ± 1.5
DTEA	− 4.34 ± 1.0	− 1.53 ± 1.1	0	0.22 ± 1.0	1.90 ± 1.3
Global (XoYoR)	− 4.57 ± 0.8	− 1.76 ± 1.1	− 0.22 ± 1.0	0	1.67 ± 1.5
KPP	− 6.24 ± 1.5	− 3.43 ± 1.5	− 1.90 ± 1.3	− 1.67 ± 1.5	0

WPI, the hybrid method yields the best segmentation result, followed by JSEG, DTEA, Global thresholding and KPP methods.

Table 3 presents the degree of superiority of each method over other methods, e.g. the second row and third column shows that hybrid method overcomes JSEG (+2.8) such that we can be 95% confident that the true difference is between 1.80 and 3.80.The fourth row and third column shows that DTEA is defeated by JSEG (− 1.53) with a true mean difference between 0.43 and 2.63.

Conclusion

This paper presents a novel approach for objective evaluation of border detection methods in dermoscopy images. In order to provide evaluation metrics that are meaningful in the context of melanoma application, we introduce specific weightings into standard metrics of sensitivity, specificity, accuracy, border error, similarity and precision. Moreover, a comprehensive metric, WPI, is proposed to facilitate comparison between different methods. The proposed WPI has also been used for the optimization of the recently proposed hybrid border detection method. The effectiveness of the proposed evaluation approach is demonstrated by applying five recent border detection methods on a set of 55 high-resolution dermoscopy images using the union of four sets of dermatologist-drawn borders as the ground truth. It is also shown that the weighted metrics, which are defined to reflect the dermatologists' perspectives, show a higher degree of agreement between automatic and manual borders, compared with standard metrics.

Acknowledgment

The assistance of Dr. George Varigos, Dr. Gayle Ross, Dr. Edward Upjohn, Dr. Alana Tuxen in obtaining the manual borders and Mr. Purves Michael and Mrs. Amanda Rebbechi in taking the dermoscopy images, all from Royal Melbourne Hospital, and Dr. Sue Finch from the Department of Mathematics and Statistics, University of Melbourne, Australia, is gratefully acknowledged. This research is supported by NICTA Victoria Research Laboratory, Australia.

References

1. "Australia skin cancer facts and figures," Available at http://www.cancer.org.au/ (accessed 1 September 2009).
2. Argenziano G, Soyer HP, Chimenti S et al. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. J Am Acad Dermatol 2003; 48: 679–693.
3. Henning JS, Dusza SW, Wang SQ, Marghoob AA, Rabinovitz HS, Polsky D, Kopf AW. The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. J Am Acad Dermatol 2007; 56: 45–52.
4. Braun RP, Rabinovitz HS, Oliviero M, Kopf AW, Saurat JH. Dermoscopy of pigmented lesions. J Am Acad Dermatol 2005; 52: 109–121.
5. Celebi ME, Iyatomi H, Schaefer G, Stoecker WV. Lesion border detection in dermoscopy images. Comput Med Imaging Graphics 2009; 33: 148–153.
6. Hintz-Madsen M, Hansen LK, Larsen J, Drzewiecki KT. A probabilistic neural network framework for the detection of malignant melanoma. Artificial Neural Networks In Cancer Diagnosis. R.N.G. Naguib and G.V. Sherbet, CRC Press 2001: 141–183.

7. Iyatomi H, Oka H, Saito M et al. Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system. Melanoma Res 2006; 16: 183–190.
8. Melli R, Grana C, Cucchiara R. Comparison of color clustering algorithms for segmentation of dermatological images. SPIE Med Imaging 2006; 6144: 3S1–3S9.
9. Schmid P. Segmentation of digitized dermatoscopic images by two-dimensional color clustering. IEEE Trans Med Imaging 1999; 18: 164–171.
10. Celebi ME, Kingravi HA, Iyatomi H et al. Border detection in dermoscopy images using statistical region merging. Skin Res Technol 2008; 14: 347–353.
11. Celebi ME, Aslandogan YA, Stoecker WV, Iyatomi H, Oka H, Chen X. Unsupervised border detection in dermoscopy images. Skin Res Technol 2007; 13: 454–462.
12. Zhou H, Chen M, Zou L, Gass R, Ferris L, Drogowski L, Rehg JM. ''Spatially constrained segmentation of dermoscopy images.'' *5th IEEE International Symposium on Biomedical Imaging*, Paris, France, 2008, pp. 800–803.
13. Garnavi R, Aldeen M, Celebi ME, Bhuiyan A, Dolianitis C, Varigos G. ''Skin lesion segmentation using color channel optimization and clustering-based histogram thresholding.'' *International Conference on Machine Vision, Image Processing, and Pattern Analysis (MVIPPA09), World Academy of Science, Engineering and Technology*, Bangkok, Thailand, vol. 60, 2009, pp. 549–557.
14. Garnavi R, Aldeen M, Celebis ME, Finch S, Varigos G. ''Border detection in dermoscopy images using hybrid thresholding on optimized color channels,'' *Computerized Medical Imaging and Graphics, Special Issue: Skin Cancer Imaging*, in press.
15. Guillod J, Schmid-Saugeon P, Guggisberg D, Cerottini JP, Braun R, Krischer J, Saurat JH, Kunt M. Validation of segmentation techniques for digital dermoscopy. Skin Res Technol 2002; 8: 240–249.
16. Celebi ME, Schaefer G, Iyatomi H, Stoecker WV, Malters JM, Grichnik JM. An improved objective evaluation measure for border detection in dermoscopy images. Skin Res Technol 2009; 15: 444–450.
17. Lee T, Ng V, Gallagher R, Coldman A, McLean D. Dullrazor: a software approach to hair removal from images. Comput Biol Med 1997; 27: 533–543.
18. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 1979; 9: 62–66.

Address:
Rahil Garnavi
NICTA Victoria Research Laboratory
Department of Electrical and Electronic Engineering
University of Melbourne
Parkville, Vic. 3010
Australia
Tel: +61 3 83 44 0377
Fax: +61 43 48 62608
e-mails: r.garnavi@ee.unimelb.edu.au, rahil.g@gmail.com