Localization of Lesions in Dermoscopy Images Using Ensembles of Thresholding Methods

M. Emre Celebi[1,⋆], Hitoshi Iyatomi[2], Gerald Schaefer[3], and William V. Stoecker[4]

[1] Department of Computer Science, Louisiana State University, Shreveport, LA, USA
ecelebi@lsus.edu
[2] Department of Electrical Informatics, Hosei University, Tokyo, Japan
[3] School of Engineering and Applied Science, Aston University, Birmingham, UK
[4] Stoecker & Associates, Rolla, MO, USA

Abstract. Dermoscopy is one of the major imaging modalities used in the diagnosis of melanoma and other pigmented skin lesions. Due to the difficulty and subjectivity of human interpretation, automated analysis of dermoscopy images has become an important research area. Border detection is often the first step in this analysis. In this article, we present an approximate lesion localization method that serves as a preprocessing step for detecting borders in dermoscopy images. In this method, first the black frame around the image is removed using an iterative algorithm. The approximate location of the lesion is then determined using an ensemble of thresholding algorithms. Experiments on a large set of images demonstrate that the presented method achieves both fast and accurate localization of lesions in dermoscopy images.

1 Introduction

Malignant melanoma, the most deadly form of skin cancer, is one of the most rapidly increasing cancers in the world, with an estimated incidence of 62,480 and an estimated total of 8,420 deaths in the United States in 2008 alone [1]. Early diagnosis is particularly important since melanoma can be cured with a simple excision if detected early.

Dermoscopy, also known as epiluminescence microscopy, has become one of the most important tools in the diagnosis of melanoma and other pigmented skin lesions. This non-invasive skin imaging technique involves optical magnification, which makes subsurface structures more easily visible when compared to conventional clinical images [2]. This in turn reduces screening errors and provides greater differentiation between difficult lesions such as pigmented Spitz nevi and small, clinically equivocal lesions [3]. However, it has also been demonstrated that dermoscopy may actually lower the diagnostic accuracy in the hands of

⋆ This work was supported by grants from the Louisiana Board of Regents (LEQSF2008-11-RD-A-12) and the Ministry of Education, Culture, Science, and Technology of Japan (Grant-in-Aid for Scientific Research C, 20591461, 2008-2010).

T. Wada, F. Huang, and S. Lin (Eds.): PSIVT 2009, LNCS 5414, pp. 1094–1103, 2009.

inexperienced dermatologists [4]. Therefore, in order to minimize the diagnostic errors that result from the difficulty and subjectivity of visual interpretation, the development of computerized image analysis techniques is of paramount importance [5,6].

Automated border detection is often the first step in the automated analysis of dermoscopy images [7,8,9]. It is crucial for the image analysis for two main reasons. First, the border structure provides important information for accurate diagnosis, as many clinical features, such as asymmetry, border irregularity, and abrupt border cutoff, are calculated directly from the border. Second, the extraction of other important clinical features such as atypical pigment networks, globules, and blue-white areas, critically depends on the accuracy of border detection.

A number of methods have been developed for preprocessing dermoscopy images. Most of these focused on the removal of artifacts such as hairs and bubbles. Of the studies dealing with hair removal, Lee *et al.* [10] approached the problem using mathematical morphology. Fleming *et al.* [5] applied curvilinear structure detection with various constraints followed by gap filling. A method for bubble removal was introduced in [5], where the authors utilized a morphological top-hat operator followed by a radial search procedure.

2 Materials and Methods

2.1 Black Frame Removal

Dermoscopy images often contain black frames that are introduced during the digitization process. These need to be removed because they might interfere with the subsequent lesion localization procedure. In order to determine the darkness of a pixel with (R, G, B) coordinates, the lightness component of the HSL color space is utilized. A pixel is considered to be black if its lightness value is less than 20. Using this criterion, the image is scanned row-by-row starting from the top. A particular row is labeled as part of the black frame if it contains 60% black pixels. The top-to-bottom scan terminates when a row that contains less than the threshold percentage of pixels is encountered. The same scanning procedure is repeated for the other three main directions.

2.2 Approximate Lesion Localization

Although dermoscopy images can be quite large, the actual lesion often occupies a relatively small area. Therefore, if we can determine the approximate location of the lesion, the border detection algorithm can focus on this region rather than the whole image. An accurate bounding box (the smallest axis-aligned rectangular box that encloses the lesion) might be useful for various reasons: (i) it provides an estimate of the lesion size (certain image segmentation algorithms such as region growing and morphological flooding can use the size of the region as a termination criterion), (ii) it might improve the border detection accuracy

since the procedure is focused on a region that is guaranteed to contain the lesion, (iii) it speeds up the border detection since the procedure is performed on a region that is often smaller than the whole image, (iv) its surrounding might be utilized in the estimation of the background skin color, which is useful for various operations including the elimination of spurious regions that are discovered during the border detection procedure [9] and the extraction of dermoscopic features such as blotches [11] and blue-white areas [12].

In many dermoscopic images, the lesion can be roughly separated from the background skin using a grayscale thresholding method applied to the blue channel [7,8]. While there are a number of thresholding methods that perform well in general, the effectiveness of a method strongly depends on the statistical characteristics of the image [13]. Fig. 1 illustrates this phenomenon[1]. Here, methods 1(d), 1(e), and 1(g) perform quite well. In contrast, methods 1(c) and 1(h) underestimate the optimal threshold, whereas method 1(f) overestimates the optimal threshold. Although method 1(c) is the most popular thresholding algorithm in the literature, for this particular image, it performs the second worst.

A possible approach to overcome this problem is to fuse the results provided by an ensemble of thresholding algorithms. In this way, it is possible to exploit the peculiarities of the participating thresholding algorithms synergistically, thus arriving at more robust final decisions than is possible with a single thresholding algorithm. We note that the goal of the fusion is not to outperform the individual thresholding algorithms, but to obtain accuracies comparable to that of the best thresholding algorithm independently of the image characteristics. In this study, we used the threshold fusion method proposed by Melgani [13], which we describe briefly in the following.

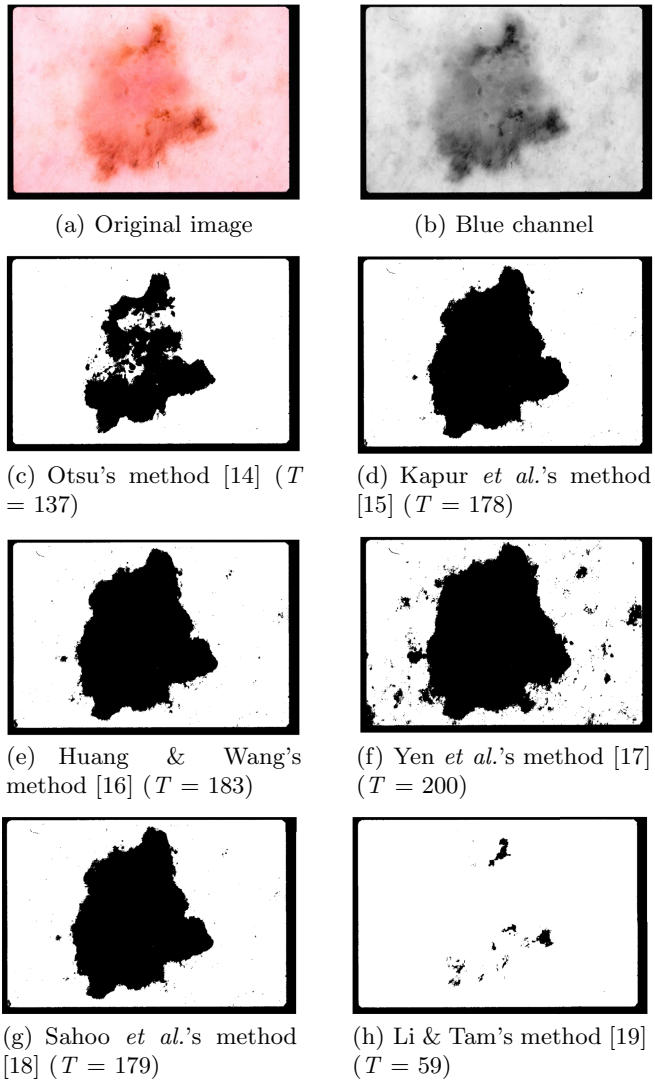
Let $X = \{x_{mn} : m = 0, 1, \ldots, M-1,\ n = 0, 1, \ldots, N-1\}$ be the original scalar $M \times N$ image with L possible gray levels ($x_{mn} \in \{0, 1, \ldots, L-1\}$) and $Y = \{y_{mn} : m = 0, 1, \ldots, M-1,\ n = 0, 1, \ldots, N-1\}$ be the binary output of the threshold fusion. Consider an ensemble of P thresholding algorithms. Let T_i and A_i ($i = 1, 2, \ldots, P$) be the threshold value and the output binary image associated with the i-th algorithm of the ensemble, respectively. Within a Markov Random Field (MRF) framework the fusion problem can be formulated as an energy minimization task. Accordingly, the local energy function U_{mn} to be minimized for the pixel (m, n) can be written as follows:

$$U_{mn} = \beta_{SP} \cdot U_{SP}\left[y_{mn}, Y^S(m,n)\right] + \sum_{i=1}^{P} \beta_i \cdot U_{II}\left[y_{mn}, A_i^S(m,n)\right] \qquad (1)$$

where S is a predefined neighborhood system associated with pixel (m, n), $U_{SP}(\cdot)$ and $U_{II}(\cdot)$ refer to the spatial and inter-image energy functions, respectively, whereas β_{SP} and β_i ($i = 1, 2, \ldots, P$) represent the spatial and inter-image parameters, respectively. The spatial energy function can be expressed as:

$$U_{SP}\left[y_{mn}, Y^S(m,n)\right] = -\sum_{y_{pq} \in Y^S(m,n)} I\left(y_{mn}, y_{pq}\right) \qquad (2)$$

[1] The frame of this image is left intact for visualization purposes.



(a) Original image (b) Blue channel

(c) Otsu's method [14] (T = 137) (d) Kapur *et al.*'s method [15] (T = 178)

(e) Huang & Wang's method [16] (T = 183) (f) Yen *et al.*'s method [17] (T = 200)

(g) Sahoo *et al.*'s method [18] (T = 179) (h) Li & Tam's method [19] (T = 59)

Fig. 1. Comparison of various thresholding methods (T: threshold)

where $I(.,.)$ is the indicator function defined as:

$$I(y_{mn}, y_{pq}) = \begin{cases} 1 & \text{if } y_{mn} = y_{pq} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The inter-image energy function is defined as:

$$U_{II}\left[y_{mn}, A_i^S(m,n)\right] = - \sum_{A_i(p,q) \in A_i^S(m,n)} \alpha^i(x_{pq}) \cdot I\left[y_{mn}, A_i(p,q)\right] \tag{4}$$

where $\alpha^i(\cdot)$ is a weight function given by:

$$\alpha^i(x_{mn}) = 1 - \exp\left(-\gamma \left| x_{mn} - T_i \right|\right) \tag{5}$$

This function controls the effect of unreliable decisions at the pixel level that can be incurred by the thresholding algorithms. At the global (image) level decisions are weighed by the inter-image parameters β_i $(i = 1, 2, \ldots, P)$, which are computed as follows:

$$\beta_i = \exp\left(-\gamma \left| \bar{T} - T_i \right|\right) \tag{6}$$

where $\bar{T}$ is the average threshold value:

$$\bar{T} = \frac{1}{P} \sum_{i=1}^{P} T_i \tag{7}$$

The MRF fusion strategy proposed in [13] is as follows:

1. Apply each thresholding algorithm of the ensemble to the image X to generate the set of thresholded images A_i $(i = 1, 2, \ldots, P)$
2. Initialize Y by minimizing for each pixel (m, n) the local energy function U_{mn} defined in Eq. 1 without the spatial energy term i.e., by setting $\beta_{SP} = 0$.
3. Update Y by minimizing for each pixel (m, n) the local energy function U_{mn} defined in Eq. 1 including the spatial energy term i.e., by setting $\beta_{SP} \neq 0$.
4. Repeat step 3 K_{max} times or until the number of different labels in Y computed over the last two iterations becomes very small.

In our preliminary experiments, we observed that, besides being computationally demanding, the iterative part (step 3) of the fusion algorithm makes only marginal contribution to the quality of the results. Therefore, in this study, we considered only the first two steps. The γ parameter was set to the recommended value of 0.1 [13]. For computational reasons, α (Eq. 5) and β (Eq. 6) values were precalculated and the neighborhood system S was chosen as a 3×3 square window.

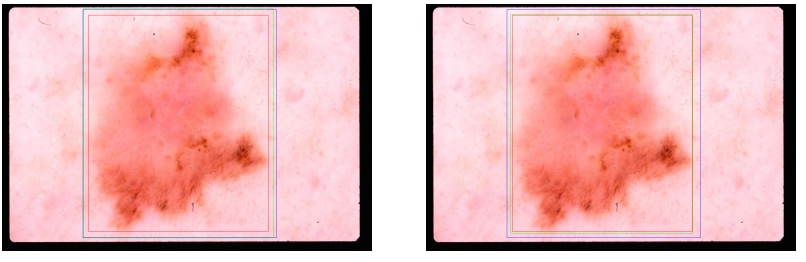
The most important performance factor in the fusion algorithm seems to be the choice of the thresholding algorithms. We considered six popular thresholding algorithms to construct the ensemble: Otsu's [14], Kapur *et al.*'s [15], Huang & Wang's [16], Yen *et al.*'s [17] , Sahoo *et al.*'s [18], and Li & Tam's [19] methods. In order to determine the best combination, we evaluated ensembles with 3 (20 ensembles), 4 (15 ensembles), 5 (6 ensembles), and 6 (1 ensemble) methods.

Fig. 2 shows the output of two particular ensembles: Otsu-Kapur-Huang and Huang-Yen-Sahoo-Li. Note that both ensembles contain at least one method that either underestimates or overestimates the optimal threshold. It can be seen that both ensembles perform equally well, which demonstrates that failures in pathological cases might be prevented using a proper fusion strategy.

Fig. 3(a) shows the result of the ensemble Otsu-Kapur-Huang-Sahoo. Here, the blue bounding box encloses the dermatologist determined border (see

(a) Otsu-Kapur-Huang (b) Huang-Yen-Sahoo-Li

Fig. 2. Comparison of two threshold ensembles



(a) Non-adaptive expansion $P = 3$ (b) Adaptive expansion $G = 6$

Fig. 3. Comparison of the bounding box expansion methods

Section 3), whereas the red one encloses the binary output of the threshold fusion. It can be seen that the red box is completely contained inside the blue box. This was observed in many cases because the automated thresholding methods tend to find the sharpest pigment change, whereas the dermatologists choose the outmost detectable pigment. We experimented with two different expansion methods to solve this problem. The first one involves expanding the automatic box by $P\%$ in four main directions. In other words, an automatic box of size $M_B \times N_B$ is expanded by $M_B \cdot P/100$ pixels in the West and East directions and $N_B \cdot P/100$ pixels in the North and South directions. The second one involves incrementing the threshold values obtained by each algorithm in the ensemble by G gray levels. In the rest of this article, we will refer to these expansion methods as non-adaptive and adaptive, respectively. Figs. 3(a) and 3(b) show the results of these methods with the expanded box shown in green. In this particular example, the non-adaptive method performs better in bringing the automatic box closer to the manual one. In order to determine the optimal expansion amounts we evaluated $P \in \{2, 4, 6, 8\}$ and $G \in \{4, 6, 8, 10\}$.

3 Results and Discussion

The proposed method was tested on a set of 428 dermoscopy images obtained from the EDRA Interactive Atlas of Dermoscopy [2] and the Keio University

Hospital. An experienced dermatologist determined the manual borders. The bounding box error was quantified using the following formula [20]:

$$\varepsilon = \frac{\text{Area}(\mathit{AutomaticBox} \oplus \mathit{ManualBox})}{\text{Area}(\mathit{ManualBox})} \cdot 100 \tag{8}$$

where *AutomaticBox* is the binary image obtained by filling the bounding box of the fusion output, *ManualBox* is the binary image obtained by filling the bounding box of the dermatologist-determined border, $\oplus$ is the exclusive-OR operation, which essentially determines the pixels for which the *AutomaticBox* and *ManualBox* disagree, and Area(I) denotes the number of pixels in the binary image I.

We determined the optimal parameter combination for the presented approximate bounding box computation method as follows. First, the black frame removal procedure described in Section 2.1 is performed on each image in the data set. The lesion bounding box is then computed using the fusion method described in Section 2.2 with one of the 42 ensembles. Finally, the approximate bounding box is expanded using either the non-adaptive method with $P \in \{2, 4, 6, 8\}$ or the adaptive method with $G \in \{4, 6, 8, 10\}$. Table 1 shows various statistics associated with the four most accurate ensembles for each expansion method. The last two columns refer to the mean and standard deviation values, respectively for the percentage image size reduction, i.e. $\frac{\text{Area}(\mathit{AutomaticBox})}{M \cdot N} \cdot 100$, provided by the bounding box computation. The following observations are in order: (i) both expansion methods reduce the mean bounding box error, (ii) the lowest mean errors were obtained using the ensemble Otsu-Kapur-Huang-Sahoo, (iii) the non-adaptive expansion method was more effective than the adaptive one, (iv) the computation of the bounding box reduced the original image size by about 260%.

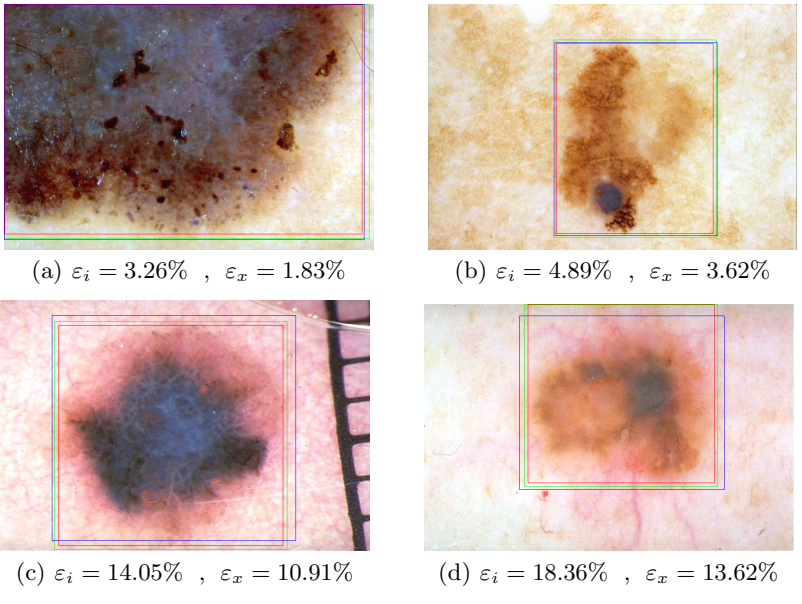
The adaptive method was less effective than the non-adaptive one probably because the former often expands the approximate box by unpredictable amounts: either too little (as in Fig. 3(b)) or too much depending on the shape of the histogram and the value of the G parameter. In contrast, the latter always expands the approximate box by an amount specified by the P parameter.

Table 1. Ensemble statistics (μ: mean, σ: std. dev., ε_i: initial box error, ε_x: expanded box error)

Ensemble	**Expansion Method**	μ_{ε_i}	σ_{ε_i}	μ_{ε_x}	σ_{ε_x}	μ_s	σ_s
Otsu-Kapur-Huang-Sahoo	Non-adaptive ($P = 2$)	10.25	8.10	7.58	8.13	268.31	185.64
Otsu-Huang-Yen-Li	Non-adaptive ($P = 4$)	11.92	7.59	7.89	6.30	260.55	183.85
Otsu-Huang-Sahoo-Li	Non-adaptive ($P = 4$)	11.98	7.62	7.90	6.20	260.95	184.14
Otsu-Huang-Sahoo	Non-adaptive ($P = 2$)	11.14	7.17	7.91	6.71	273.84	195.69
Otsu-Kapur-Huang-Sahoo	Adaptive ($G = 6$)	10.25	8.10	9.27	7.68	276.92	192.14
Kapur-Huang-Sahoo-Li	Adaptive ($G = 8$)	10.98	7.66	9.43	7.69	279.03	194.42
Otsu-Kapur-Huang-Sahoo	Adaptive ($G = 4$)	10.25	8.10	9.44	7.56	279.98	194.26
Kapur-Huang-Sahoo-Li	Adaptive ($G = 6$)	10.98	7.66	9.67	7.58	282.09	196.58

Table 2. Individual statistics (μ: mean, σ: std. dev., ε_i: initial box error, ε_x: expanded box error)

Thresholding Method	Expansion Method	μ_{ε_i}	σ_{ε_i}	μ_{ε_x}	σ_{ε_x}	μ_s	σ_s
Otsu	Non-adaptive ($P = 2$)	12.05	9.10	9.00	8.95	275.07	199.28
Kapur	Non-adaptive ($P = 2$)	12.87	16.86	12.68	17.56	261.95	197.94
Huang	Non-adaptive ($P = 2$)	20.31	67.97	17.17	69.76	269.59	190.09
Yen	Non-adaptive ($P = 2$)	14.98	27.12	15.74	27.74	255.61	250.53
Sahoo	Non-adaptive ($P = 2$)	13.43	24.60	13.37	25.19	254.43	184.36
Li	Non-adaptive ($P = 2$)	15.12	9.65	11.06	9.07	293.54	215.80
Otsu	Non-adaptive ($P = 4$)	12.05	9.10	9.10	9.14	256.86	182.82
Kapur	Non-adaptive ($P = 4$)	12.87	16.86	15.54	18.61	245.36	183.78
Huang	Non-adaptive ($P = 4$)	20.31	67.97	16.83	70.69	251.99	174.44
Yen	Non-adaptive ($P = 4$)	14.98	27.12	19.32	28.49	239.46	230.91
Sahoo	Non-adaptive ($P = 4$)	13.43	24.60	16.43	25.98	238.32	170.38
Li	Non-adaptive ($P = 4$)	15.12	9.65	9.41	7.99	273.93	198.61



(a) $\varepsilon_i = 3.26\%$, $\varepsilon_x = 1.83\%$ (b) $\varepsilon_i = 4.89\%$, $\varepsilon_x = 3.62\%$

(c) $\varepsilon_i = 14.05\%$, $\varepsilon_x = 10.91\%$ (d) $\varepsilon_i = 18.36\%$, $\varepsilon_x = 13.62\%$

Fig. 4. Sample results (ε_i: initial box error, ε_x: expanded box error)

Table 2 shows the statistics for the individual thresholding methods. Note that, due to space limitations, we report only the results of the non-adaptive expansion method (as in the ensemble case, the adaptive method has inferior performance). It can be seen that, in most configurations, the individual methods obtain significantly higher mean errors than the best ensemble methods, i.e. the first four rows of Table 1. This is because, as explained in Section 2.2, the individual methods are more prone to catastrophic failures when given pathological

input images. The high standard deviation values also support this explanation. Only the performance of Otsu (with $P = 2, 4$) and Li *et al.*'s (with $P = 4$) methods is close to the performance of the ensembles. However, as mentioned in Section 2.2, the goal of fusion is not to outperform the individual thresholding algorithms, but to obtain accuracies comparable to that of the best thresholding algorithm independently of the image characteristics.

Fig. 4 shows sample bounding box computation results obtained using the ensemble Otsu-Kapur-Huang-Sahoo with $P = 2$. It can be seen that the presented method determines an accurate bounding box even for lesions with fuzzy borders.

4 Conclusions

In this paper, an automated method for approximate lesion localization in dermoscopy images is presented. The method is comprised of three main phases: black frame removal, initial bounding box computation using an ensemble of thresholding algorithms, and expansion of the initial bounding box. The execution time of the method is about 0.15 seconds for a typical image of size 768×512 pixels on an Intel Pentium D 2.66Ghz computer.

The presented method may not perform well on images with significant amount of hair or bubbles since these elements alter the histogram, which in turn results in biased threshold computations. Future work will be directed towards testing the utility of this method in a border detection study. The implementation of the threshold fusion method will be made publicly available as part of the Fourier image processing and analysis library, which can be downloaded from http://sourceforge.net/projects/fourier-ipal

References

1. Jemal, A., Siegel, R., Ward, E., et al.: Cancer Statistics. CA: A Cancer Journal for Clinicians 2008 58(2), 71–96 (2008)
2. Argenziano, G., Soyer, H.P., De Giorgi, V., et al.: Dermoscopy: A Tutorial. EDRA Medical Publishing & New Media, Milan (2002)
3. Steiner, K., Binder, M., Schemper, M., et al.: Statistical Evaluation of Epiluminescence Dermoscopy Criteria for Melanocytic Pigmented Lesions. Journal of American Academy of Dermatology 29(4), 581–588 (1993)
4. Binder, M., Schwarz, M., Winkler, A., et al.: Epiluminescence Microscopy. A Useful Tool for the Diagnosis of Pigmented Skin Lesions for Formally Trained Dermatologists. Archives of Dermatology 131(3), 286–291 (1995)
5. Fleming, M.G., Steger, C., Zhang, J., et al.: Techniques for a Structural Analysis of Dermatoscopic Imagery. Computerized Medical Imaging and Graphics 22(5), 375–389 (1998)
6. Celebi, M.E., Kingravi, H.A., Uddin, B., et al.: A Methodological Approach to the Classification of Dermoscopy Images. Computerized Medical Imaging and Graphics 31(6), 362–373 (2007)

7. Iyatomi, H., Oka, H., Saito, M., et al.: Quantitative Assessment of Tumor Extraction from Dermoscopy Images and Evaluation of Computer-based Extraction Methods for Automatic Melanoma Diagnostic System. Melanoma Research 16(2), 183–190 (2006)
8. Celebi, M.E., Aslandogan, Y.A., Stoecker, W.V., et al.: Unsupervised Border Detection in Dermoscopy Images. Skin Research and Technology 13(4), 454–462 (2007)
9. Celebi, M.E., Kingravi, H.A., Iyatomi, H., et al.: Border Detection in Dermoscopy Images Using Statistical Region Merging. Skin Research and Technology 14(3), 347–353 (2008)
10. Lee, T.K., Ng, V., Gallagher, R., et al.: Dullrazor: A Software Approach to Hair Removal from Images. Computers in Biology and Medicine 27(6), 533–543 (1997)
11. Stoecker, W.V., Gupta, K., Stanley, R.J., et al.: Detection of Asymmetric Blotches in Dermoscopy Images of Malignant Melanoma Using Relative Color. Skin Research and Technology 11(3), 179–184 (2005)
12. Celebi, M.E., Iyatomi, H., Stoecker, W.V., et al.: Automatic Detection of Blue-White Veil and Related Structures in Dermoscopy Images. Computerized Medical Imaging and Graphics 32(8) (to appear, 2008)
13. Melgani, F.: Robust Image Binarization with Ensembles of Thresholding Algorithms. Journal of Electronic Imaging 15(2), 023010, 11 pages (2006)
14. Otsu, N.: A Threshold Selection Method from Gray Level Histograms. IEEE Trans. on Systems, Man and Cybernetics 9(1), 62–66 (1979)
15. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram. Graphical Models and Image Processing 29(3), 273–285 (1985)
16. Huang, L.-K., Wang, M.-J.J.: Image Thresholding by Minimizing the Measures of Fuzziness. Pattern Recognition 28(1), 41–51 (1995)
17. Yen, J.C., Chang, F.J., Chang, S.: A New Criterion for Automatic Multilevel Thresholding. IEEE Trans. on Image Processing 4(3), 370–378 (1995)
18. Sahoo, P.K., Wilkins, C., Yeager, J.: Threshold Selection Using Renyi's Entropy. Pattern Recognition 30(1), 71–84 (1997)
19. Li, C.H., Tam, P.K.S.: An Iterative Algorithm for Minimum Cross Entropy Thresholding. Pattern Recognition Letters 18(8), 771–776 (1998)
20. Hance, G.A., Umbaugh, S.E., Moss, R.H., Stoecker, W.V.: Unsupervised Color Image Segmentation with Application to Skin Tumor Borders. IEEE Engineering in Medicine and Biology 15(1), 104–111 (1996)