Evaluation of the Menzies method potential for automatic dermoscopic image analysis

André R. S. Marcal & Teresa Mendonca
Faculdade de Ciências, Universidade do Porto, Portugal

Cátia S. P. Silva
Faculdade de Engenharia, Universidade do Porto, Portugal

Marta A. Pereira & Jorge Rozeira
Hospital Pedro Hispano, Matosinhos, Portugal

ABSTRACT: There is a considerable interest in the development of automatic image analysis systems for dermoscopic images. The standard approach usually consists of three stages: (i) image segmentation, (ii) feature extraction and selection, and (iii) lesion classification. This paper evaluates the potential of an alternative approach, based on the Menzies method. It consists on the identification of the presence of 1 or more of 6 possible color classes, indicating that the lesion should be considered a potential melanoma. The Jeffries-Matusita (JM) and Transformed Divergence (TD) separability measures were used for an experimental evaluation with 28 dermoscopic images. In the most challenging case tested, with training identified in multiple images, 8 out of 15 class pairs were found to be well separable, or 13+2 out of 21 considering the skin as an additional class.

1 INTRODUCTION

Dermoscopy (dermatoscopy or skin surface microscopy) is a non-invasive diagnostic technique for the in vivo observation of pigmented skin lesions used in dermatology. The automatic analysis of dermoscopic images is of great interest, both to provide quantitative information about a lesion, for example to support the follow up procedures by the clinician, and also as a potential stand-alone early warning tool. In the last few years a number of screening tests have been proposed for dermoscopic images, suitable for health care personal with minimum clinical training. These screening tests, such as the ABCD Rule algorithm (Marghoob & Braun 2004) and the 7-point check-list algorithm (Argenziano et al. 2011) are used to reduce the number of cases that need to be evaluated by a dermatologist.

Various attempts have been made to implement computer based systems inspired on the human based screening tests (Mendonça et al. 2007). The standard approach of these automatic dermoscopic image analysis systems usually consists of three stages: (i) image segmentation, (ii) feature extraction and selection, and (iii) lesion classification. The segmentation procedure alone is a challenging task as the various algorithms produce different segmentation results (Silveira et al. 2009). Furthermore, even when the lesion is segmented, there is still considerable work to be done in order to establish a link between the human based criteria and the features extracted by automatic computer based algorithms.

An alternative approach for the evaluation of dermoscopic images is the Menzies method, where the presence of 1 or more out of 6 color classes indicates that the lesion should be considered a potential melanoma (Menzies 2001). The implementation of an image processing system based on the Menzies method does not require the segmentation stage, nor the subsequent extraction of geometric features, avoiding some of the potential errors in those stages. The purpose of this work is to investigate the applicability of the Menzies method for the development of an image processing tool for dermoscopic image analysis.

2 METHODS

2.1 *Separability measurements*

Dermoscopic images are usually RGB color images, having 3 independent channels (the Red, Green and Blue color channels). An image pixel, or observation, is characterized by a vector $x = [x_1, x_2, ..., x_n]^T$, where xi is the grey level intensity of image channel i. Thus, for a typical dermoscopic RGB color image (n=3), there are only 3 features. Each color class corresponds to a specific location and volume in the 3D feature space. The regions of the feature space associated with each color class would ideally be non-overlapping and well spread. However, this is rarely the case in real life problems, as the data from the various channels tend to be highly correlated, which results in the color class regions being mostly locat-

ed along the feature space diagonal. It is therefore important to quantify the separability between the various classes in order to evaluate the potential applicability of the Menzies method for an automatic image processing system.

Divergence is a separability measure for a pair of probability distributions that has its basis in their degree of overlap. It is defined in terms of the likelihood ratio $L_{ij}(x) = p(x|\omega_i) / p(x|\omega_j)$, where $p(x|\omega_i)$ and $p(x|\omega_j)$ are the values of the probability distributions of the spectral classes i and j, for a feature vector x (Richards & Jia, 2006). For perfectly separable classes, $L_{ij}(x)=0$ or $L_{ij}(x)=\infty$ for all values of x.

It is worth choosing the logarithm of the likelihood ratio (L') by means of which the divergence of the pair of class distribution is defined as $d_{ij} = E\{L'_{ij}|\omega_i\} + E\{L'_{ij}|\omega_j\}$, where E is the expectation operator (Richards & Jia 2006). For spectral classes modeled by multi-dimensional normal distributions, with means μ_i, and μ_j, and covariances Σ_i and Σ_j, it can be shown that:

$$d_{i,j} = \frac{1}{2} tr\{(\Sigma_i - \Sigma_j)(\Sigma_i^{-1} - \Sigma_j^{-1})\} + tr\{(\Sigma_i^{-1} - \Sigma_j^{-1})(m_i - m_j)(m_i - m_j)^T\} \quad (1)$$

where tr{} is the trace of the subject matrix (Richards & Jia 2006).

The Bhattacharyya distance $D_B(i,j)$ between two classes can be calculated from the variance and mean of each class (Reyes-Aldasoro & Bhalerao 2006), in the following way:

$$D_B(i,j) = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln\left\{\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}\right\} \quad (2)$$

The Jeffries-Matusita (JM) distance J_{ij} is used to assess how well two classes are separated. The JM distance between a pair of probability distributions (spectral classes) is defined as:

$$J_{i,j} = \int_x \left\{\sqrt{p(x \mid \omega i)} - \sqrt{p(x \mid \omega j)}\right\}^2 dx \quad (3)$$

which is seen to be a measure of the average distance between density functions (Richards & Jia 2006). For normally distributed classes it becomes:

$$J_{i,j} = 2(1 - e^{-D_B}) \quad (4)$$

where D_B is the Bhattacharyya distance (Equation 2). The presence of the exponential factor in Equation 4 gives an exponentially decreasing weight to increasing separations between spectral classes. If plotted as a function of distance between class means it shows a saturating behavior, not unlike that expected for the probability of correct classification. It is asymptotic to 2, so that for a JM distance of 2, the signatures can be totally separable (with 100% accuracy). Generally, the classes can be considered separable for JM values above 1.8 or, preferably, above 1.9. JM values bellow 1.8 indicate the possible confusion between the class pair in the classification process (Richards & Jia 2006).

An useful modification of divergence (Eq.1) becomes apparent by noting the algebraic similarity of d_{ij} to the parameter D_B, used in the JM distance. Since both involve terms which are functions of the covariance alone, and terms which appear as normalised distances between class means, it is possible to make use of a heuristic Transformed Divergence (TD) measure of the form (Swain & Davis 1978):

$$d_{i,j}{}^T = 2(1 - e^{-\frac{d_{i,j}}{8}}) \quad (5)$$

Because of its exponential character, TD will have saturating behavior with increasing class separation, as does the JM distance, and yet it is computationally more economical. It is worth noting that for the computational implementation of d_{ij} and $D_B(i,j)$, and consequently for the JM and TD distances, the data must be Gaussian.

2.2 *Image calibration*

Dermoscopic images are acquired under controlled illumination and should thus produce accurate color images. However, this is not always true in practice and a dermatologist is expected to make the same diagnosis on a particular case even if the image is acquired in different imaging conditions (Iyatomi et al. 2011). Furthermore, the patient skin (background) varies, which results in a different color perception of the lesion by the clinician. Although color calibration in dermoscopy is recognized as important, it is still an open issue (Iyatomi et al. 2011).

Two alternative calibration procedures were considered here. The assumption is that the background (skin) should be normalized to a reference. Sample areas for skin ($\mathbf{m}^{skin}$) and lesion ($\mathbf{m}^{lesion}$) are considered, with $\mathbf{m}$ being a 3-dimensional vector with the average RGB values for the sample. Reference vectors ($\mathbf{r}^{skin}$ and $\mathbf{r}^{lesion}$) are also considered.

Calibration method I uses only the skin intensity. The original RGB values are multiplied by a factor α, computed as $\alpha = \| \mathbf{r}^{skin} \| / \| \mathbf{m}^{skin} \|$.

Calibration method II applies different multiplication factors (β_j) for the 3 color components of the RGB image. The coefficients (β_j) are computed as $\beta_j = (\mathbf{r}_j^{skin} - \mathbf{r}_j^{lesion}) / (\mathbf{m}_j^{skin} - \mathbf{m}_j^{lesion})$, with j=1,2,3.

3 RESULTS

3.1 *Experimental data description*

An experimental procedure was devised to test the applicability of the Menzies method – presence of at least 1 of 6 possible colors – for an automatic image processing system. A total of 28 dermoscopic images (IMD01 to IMD28) were selected and evaluated by a

dermatologist, identifying the presence and location of the six Menzies color classes – Black (Bk), Blue Gray (BG), Dark Brown (DB), Light Brown (LB), Red (Rd) and White (Wh). The identification was done using database and segmentation tools developed in the Automatic computer-based Diagnosis system for Dermoscopy Image (ADDI) project (Amorim et al. 2011, Ferreira et al. 2011). As an example, Figure 1 shows the original dermoscopic image (IMD19) and the segmented regions associated with the color classes Dark Brown (DB) and Blue Gray (BG), both identified as present in this image. Most test images have a single color (13) or two colors (12), with only 3 out of 28 images having three Menzies colors present in a single lesion. A summary of the color classes present in the various test images is available in Table 1.

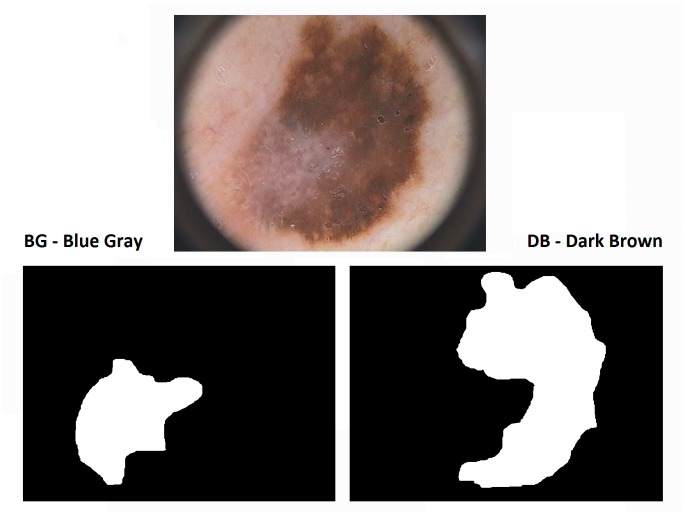

Figure 1. Example of a dermoscopic image (IMD19) and the medical evaluation (segmentation) for color classes BG – Blue Gray (left) and DB – Dark Brown (right).

Table 1. Menzies colors present in the test image dataset.

Colors	Wh	Re	LB	DB	BG	Bk	Images (IMDxx)
			x				05, 08, 10, 14,
			x				23, 24, 25
1				x			02, 07, 15, 16, 21
						x	18
			x	x			01, 03, 04, 06
			x	x			09, 11, 22
2				x	x		12, 19, 26
		x	x				27
			x			x	13
	x			x	x		17
3	x				x	x	20
			x	x	x		28

3.2 *Single Image Training*

Initially, training areas for each class were identified in a single image. The images used were IMD17 (for white), IMD27 (red), IMD05 (light brown), IMD01 (dark brown), IMD12 (blue gray) and IMD20 (black). The 3D scatterplot and the projection in the 2D plane (Red/Blue) are presented in Figure 2. As expected, the feature space is scarcely used and there is a certain amount of overlap between some classes.

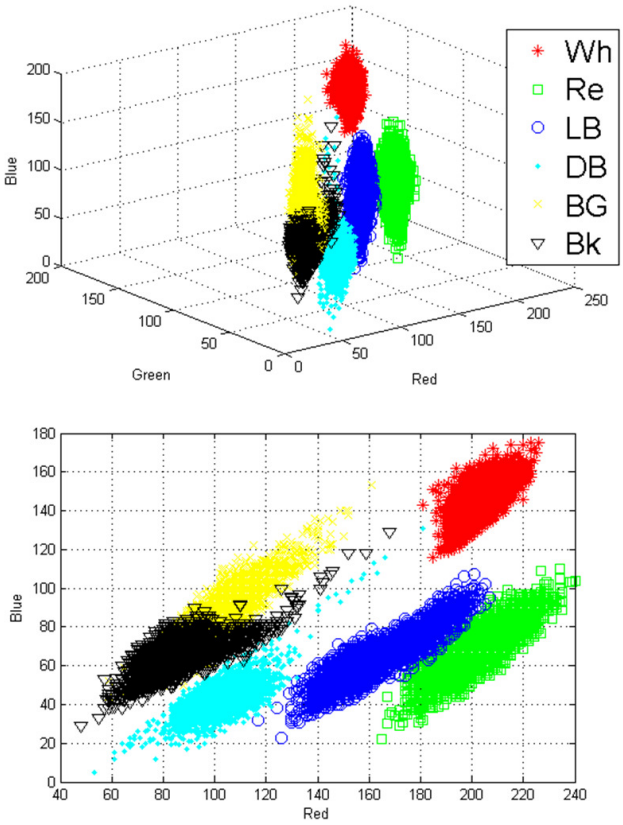

Figure 2. Scatterplots for the single image training test: on the RGB feature space (top) and a projection on the 2D Red, Blue plane (bottom).

The Kolmogorov-Smirnov (KS) test was used to determine if the samples can be considered as originating from a normal (Gaussian) distribution (Kolmogorov 1993). The KS test provides the mean (μ_i) and standard deviation (σ_i) parameters for each sample i, as well as the probability of the data to have a normal distribution (p-value). The color classes follow Gaussian distributions, as demonstrated by the p-values results for significance level $\alpha = 5\%$ presented in Table 2 (all above 0.90). Table 2 also presents the minimum, maximum, average (μ) and standard deviation (σ) values for each color class.

Considering that the color classes have a normal distribution, it is thus possible to use the JM (Eq.4) and TD (Eq.5) distances to evaluate the class separabilities. Table 3 presents the two triangular matrices (TD and JM) together, the JM distance as a top right triangular matrix and the TD as a bottom left triangular matrix. For most class pairs, both TD and JM have either a perfect separability (2.00), or very high values (above 1.9). The only major difficulty is to distinguish between color classes BG and Bk, which have low values of both TD and JM distances. According to this test, considering the RGB color features and training data used, these two classes are not distinguishable, but all other class pairs are well separable.

Table 2. Characteristics summary for the six classes in the RGB color channels for the single image training case.

	Class	Min	Max	μ	σ	p
	Wh	181	226	205	6.1	0.99
R	Re	165	240	205	10.4	0.95
E	LB	117	207	164	15.3	0.91
D	DB	53	181	102	9.4	0.96
	BG	59	161	101	9.5	0.95
	Bk	48	168	83	8.4	0.96
	Wh	133	181	156	5.8	0.99
G	Re	70	152	114	11.1	0.94
R	LB	63	145	104	14.4	0.90
E	DB	14	142	59.2	7.5	0.98
E	BG	51	147	86.0	10.6	0.95
N	Bk	38	133	68.1	6.0	0.98
	Wh	116	171	148	8.5	0.96
B	Re	22	110	70.4	10.7	0.93
L	LB	23	107	64.5	12.5	0.94
U	DB	5	131	42.9	7.9	0.95
E	BG	44	153	84.4	12.1	0.92
	Bk	29	129	64.8	6.4	0.99

Table 3. Jeffries-Matusita - JM (top right triangular matrix) and Transformed Divergence - TD (bottom left) distance values, for the single image training test.

	Wh	Re	LB	DB	BG	Bk
Wh	-	2.00	2.00	2.00	2.00	2.00
Re	2.00	-	2.00	2.00	2.00	2.00
LB	2.00	2.00	-	1.98	2.00	2.00
DB	2.00	2.00	2.00	-	1.99	1.99
BG	2.00	2.00	2.00	2.00	-	1.02
Bk	2.00	2.00	2.00	1.99	1.01	-

3.3 *Multiple Image Training*

A second test, more challenging, was performed, by identifying the training areas for each color class using all 28 test images available. As it can be seen in Table 1, the number of images used to train each color class varies considerably, from only a few (1 for Re and 2 for Wh) up to 17 (for both LB and DB). The large number of images used to train these classes (DB and LB) greatly increases the dispersion of their signatures in the RGB feature space, thus increasing the overlap between the various color classes.

The p-values from the KS test are generally lower than those for single image training. For class Bk the p-values are above 0.74 for the 3 color channels (R, G, B), while for classes Wh and Re the p-values are all above 0.92. For classes DB and BG the p-values vary between 0.54 and 0.75 for the three color channels. The worst class is class LB, with p-values of 0.56, 0.57 and 0.50 for the R, G and B color channels. The data was nevertheless still treated as having normal distribution, and the TD and JM distances were thus computed.

Table 4 presents the two triangular matrices for the multiple-image training test (the JM distance at top right and the TD at bottom left). These results indicate that 8 out of 15 classes pairs are clearly distinguishable (for JM) or 9 out of 15 (based on TD values). There is a class pair (Wh-LB) with contradictory indications from JM and TD distances (1.74 and 1.98, respectively). The remaining 6 class pairs have both low JM and TD values (below 1.8) and cannot thus be considered separable. These class pairs are: Wh-LB, Wh-BG, Re-LB, LB-DB, DB-BG, DB-Bk, BG-Bk. More details about the multiple image training evaluation, with a slightly different dataset, are available in Silva et al. (2012).

Table 4. Jeffries-Matusita - JM (top right triangular matrix) and Transformed Divergence - TD (bottom left) distance values, for the multiple-image training test.

	Wh	Re	LB	DB	BG	Bk
Wh	-	2.00	1.74	1.97	1.69	1.99
Re	2.00	-	1.48	1.97	2.00	2.00
LB	1.98	1.23	-	1.47	1.87	1.99
DB	1.95	2.00	1.21	-	1.03	1.22
BG	1.70	2.00	1.94	0.75	-	0.55
Bk	1.99	2.00	2.00	0.94	0.10	-

3.4 *Image calibration*

Another test was carried out, both for the single and the multiple image training cases, using calibrated versions of the test images. The JM and TD values, computed with the calibrated images, are presented in Table 5 for calibration method I and in Table 6 for method II. There are a few cases where the indications from JM and TD distances are contradictory, and also some cases where it was not possible to calculate TD due to the fact that the data matrix was not symmetrical.

Table 5. Jeffries-Matusita - JM (top right triangular matrix) and Transformed Divergence - TD (bottom left) distance values, for the multiple-image training test with calibration I.

	Wh	Re	LB	DB	BG	Bk
Wh	-	2.00	1.86	1.96	1.73	2.00
Re	2.00	-	1.78	1.97	2.00	2.00
LB	1.83	1.77	-	1.48	1.84	2.00
DB	2.00	2.00	1.49	-	0.78	0.96
BG	1.98	2.00	1.78	0.23	-	1.11
Bk	2.00	2.00	(*)	(*)	(*)	-

(*) Not possible to calculate

Table 6. Jeffries-Matusita - JM (top right triangular matrix) and Transformed Divergence - TD (bottom left) distance values, for the multiple-image training test with calibration II.

	Wh	Re	LB	DB	BG	Bk
Wh	-	2.00	1.92	1.93	1.65	2.00
Re	2.00	-	2.00	2.00	2.00	2.00
LB	(*)	2.00	-	1.31	1.47	1.79
DB	1.33	2.00	(*)	-	0.83	1.63
BG	1.98	2.00	0.76	(*)	-	1.57
Bk	2.00	2.00	2.00	(*)	(*)	-

(*) Not possible to calculate

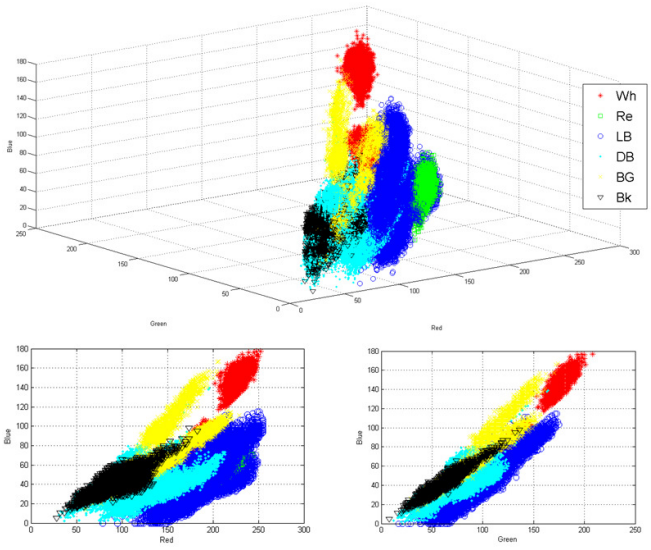

Figure 3. Scatterplots for the multiple image training test with calibration I. RGB feature space (top) and projection on 2D: Green, Blue (bottom left) and Red, Blue (bottom right) planes.

A 3D scatterplot of the multiple-image training data with calibration method I is presented in Figure 3, as well as projections in 2D planes for the Red/Blue and Green/Blue color channels. As in the single-image training case (Figure 2) the feature space is scarcely used and there is a considerable overlap between some classes.

The results presented in Tables 3-6 is summarized in Table 7, where the number of class pairs (15 in total) are grouped in well separable (JM≥1.9), moderately separable (1.8≤JM<1.9) and non-separable or undistinguishable (JM<1.8). As expected, class separability evaluation benefited slightly by performing image calibration to a reference. The best results were obtained by the calibration method I. The results based on TD values are slightly better, but as it was not always possible to compute TD, the evaluation is mostly based on JM values.

Table 7. Summary of JM results for the various scenarios tested (number of class pairs).

Test scenario	JM≥1.9	1.8≤JM<1.9	JM<1.8
Single-image training	14	-	1 (a)
Multiple-image training	7	1	7 (b)
M-image train., Calib. I	7	2	6 (c)
M-image train., Calib. II	8	-	7 (d)

(a) BG-Bk.
(b) Wh-LB, Wh-BG, Re-LB, LB-DB, DB-BG, DB-Bk, BG-Bk.
(c) Wh-BG, Re-LB, LB-DB, DB-BG, DB-Bk, BG-Bk.
(d) Wh-BG, LB-DB, LB-BG, LB-Bk, DB-BG, DB-Bk, BG-Bk.

3.5 *Including the Skin*

A final test was carried out, including an additional class – the skin. The skin can be seen as the background of the lesion in a dermoscopic image, but also as a class on its own. This additional class increases the difficulty in the classification / discrimination problem. The JM and TD values were computed for the 7 class case (21 pairs). A summary of the JM results for the various scenarios tested is presented in Table 8. The table shows the number of class pairs considered well separable in the feature space (JM≥1.9), moderately separated (1.8≤JM<1.9) and undistinguishable (JM<1.8). For the single-image training case, most class pairs are distinguishable, even with the additional skin class. However, when the training uses all 28 test images, there are 6 class pairs that cannot be separable using the RGB feature space. When the test images are calibrated to a reference (using method I), the number of distinguishable class pairs increases to 13+2 (out of 21).

Table 8. Summary of JM results for the various scenarios tested (number of class pairs).

Test scenario	JM≥1.9	1.8≤JM<1.9	JM<1.8
Single-image training	19	-	2 (a)
Multiple-image training	11	1	9
M-image train., Calib. I	13	2	6 (b)
M-image train., Calib. II	10	-	11

(a) Bk-BG, Wh-skin.
(b) Wh-BG, Re-LB, LB-DB, DB-BG, DB-Bk, BG-Bk.

4 CONCLUSIONS

An alternative approach to the standard computer based analysis of dermoscopic images is offered by the Menzies diagnosis method. The method consists on the identification of the presence of colors out of 6 possible color classes in a dermoscopic image. This method has some advantages comparing to the standard approaches, as it does not require the lesion segmentation. However, the identification of color classes in dermoscopic images is a subjective task, which poses great challenges for an automatic implementation. The purpose of this work was to evaluate the potential discrimination between the various Menzies color classes in dermoscopic RGB images.

The tests performed using the JM and TD separability metrics indicate that it is possible to identify and distinguish most color classes. In the most challenging case, where the class signature is obtained by multiple images (between 1 and 17 for each class, on a total of 28 test images), 7+1 of the 15 class pairs are distinguishable without image calibration. The result is slightly improved by using a simple calibration procedure, which normalizes the background intensity. Considering the skin as an additional class, a total of 13 class pairs are well separable and another 2 pairs are moderately separable (using calibration method I). Only 6 class pairs remain as non-separable in both cases (with or without the skin class).

Although most class pairs are distinguishable in RGB color images, the Menzies method cannot yet be used as there are considerable confusion between some color classes. Further work is thus required in order to establish more suitable features and calibration procedures. An alternative can be the use of

color models that have the color component detached from the intensity (e.g. HSI model). However, the preliminary tests carried out using the HSI color model only produced slightly improved results.

Another line of work should focus on the subjectivity in the human perception of color. In order to evaluate this aspect, a more extensive collection of images and clinical evaluation is needed, preferably with more than one medical evaluation for each dermoscopic image.

5 ACKNOWLEDGEMENTS

This work was done as part of the ADDI project (http://www2.fc.up.pt/addi/). The authors would like to thank Bárbara Amorim for her support in the acquisition of training data.

6 REFERENCES

Amorim, B., Marçal, A., Mendonça, T., Marques, J.S. & Rozeira, J. 2011. Database implementation for clinical and computer assisted diagnosis of dermoscopic images. In João Manuel R.S. Tavares & R.M. Natal Jorge (eds), *VipIMAGE; Proc. intern. Symp, Olhão, Portugal, 12-14 October 2011*. Taylor & Francis.

Argenziano, G., Catricala, C., Ardigo, A., Buccini, P., De Simone, P., Eibenschutz, L., Ferrari, A., Mariani, G., Silip, V., Sperduti, I. & Zalaudek, I. 2011. Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology* 164(4): 785-790.

Ferreira, P.M., Mendonça, T., Rocha, P. & Rozeira, J. 2011. A new interface for manual segmentation of dermoscopic images. In João Manuel R.S. Tavares & R.M. Natal Jorge (eds), *VipIMAGE; Proc. intern. Symp, Olhão, Portugal, 12-14 October 2011*. Taylor & Francis.

Iyatomi, H., Celebi, M.E., Schaefer, G. & Tanaka, M. 2011. Automated color calibration method for dermoscopy images. *Computerized Medical Imaging and Graphics* 35 89-98.

Kolmogorov, A.N. 1993. Sulla determinazione empirica di una legge di distributione. *Giornale dell'Istituto Italiano degli Attuari* 4: 83-91.

Marghoob, A. & Braun, R. 2004. *Atlas of dermoscopy*. Taylor and Francis.

Mendonça, T., Marçal, A.R.S., Vieira, A., Lacerda, L., Caridade, C. & Rozeira, J. 2007. Automatic analysis of dermoscopy images - A review. In João Manuel R.S. Tavares & R.M. Natal Jorge (eds), *Computational Modelling of Objects Represented in Images*: 281-287. Taylor & Francis.

Menzies, S.W. 2001. A method for the diagnosis of primary cutaneous melanoma using surface microscopy. *Dermatologic Clinics* 19(2): 299-305.

Reyes-Aldasoro, C.C. & Bhalerao, A. 2006. The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition* 39(5) 812-826.

Richards, J.A. & Jia, X. 2006. *Remote Sensing Digital Image Analysis: An Introduction*, 4th edition. Springer.

Silva, C.S.P., Marcal, A.R.S., Pereira, M.A., Mendonça, T. & Rozeira, J, 2012. Separability analysis of color classes on dermoscopic images. *Lecture Notes on Computer Science* (in press).

Silveira, M., Nascimento, J.C., Marques, J.S., Marçal, A.R.S., Mendonça, T., Yamauchi, S., Maeda, J. & Rozeira, J, 2009. Comparison of Segmentation Methods for Melanoma Diagnosis in Dermoscopy Images. *IEEE Journal of Selected Topics in Signal Processing* 3(1): 35-45.

Swain, P.H. & Davis, S.M., 1978. Remote Sensing: The Quantitative Approach. New York: McGraw-Hill.