Pattern Recognition Letters 32 (2011) 2187–2196

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions

Germán Capdehourat [a], Andrés Corez [a], Anabella Bazzano [b], Rodrigo Alonso [a], Pablo Musé [a,*]

[a] IIE, Facultad de Ingeniería, Universidad de la República, Uruguay
[b] Cátedra de Dermatología, Hospital de Clínicas, Facultad de Medicina, Universidad de la República, Uruguay

ARTICLE INFO

Article history:
Available online 26 June 2011

Keywords:
Melanoma
Dermoscopy
Pigmented skin lesion classification
Adaptive boosting
Support vector machines
Decision trees

ABSTRACT

In this paper we propose a machine learning approach to classify melanocytic lesions as malignant or benign, using dermoscopic images. The lesion features used in the classification framework are inspired on border, texture, color and structures used in popular dermoscopy algorithms performed by clinicians by visual inspection. The main weakness of dermoscopy algorithms is the selection of a set of weights and thresholds, that appear not to be robust or independent of population. The use of machine learning techniques allows to overcome this issue. The proposed method is designed and tested on an image database composed of 655 images of melanocytic lesions: 544 benign lesions and 111 malignant melanoma. After an image pre-processing stage that includes hair removal filtering, each image is automatically segmented using well known image segmentation algorithms. Then, each lesion is characterized by a feature vector that contains shape, color and texture information, as well as local and global parameters. The detection of particular dermoscopic patterns associated with melanoma is also addressed, and its inclusion in the classification framework is discussed. The learning and classification stage is performed using AdaBoost with C4.5 decision trees. For the automatically segmented database, classification delivered a specificity of 77% for a sensitivity of 90%. The same classification procedure applied to images manually segmented by an experienced dermatologist yielded a specificity of 85% for a sensitivity of 90%.

1. Introduction

The medical term *melanoma* refers to a malignant tumor developed from melanocytic cells. Melanoma generally appears *de novo*, and less frequently as the evolution of acquired benign melanocytic nevi. The strong metastatic power of this tumor leads to significantly high mortality rates. In the last decades, mainly due to sun exposure, the incidence of melanoma has dramatically increased, particularly in young white population. For instance, in North America it is now the fifth most common cancer among males and the sixth most common cancer among females; Australia, where melanoma is the fourth most common registered cancer, has the highest incidence in the world (see Soyer et al., 2007, chap. 4, and references therein). Early diagnosis and removal of thin melanoma, when the chance of metastasis is low, is the most effective strategy. If diagnosed and treated early, the mean life expectancy of individuals suffering from melanoma can be increased by at least 25 years.

The features that differentiate a common nevi from a melanoma tend to develop as tumor grows. Consequently, the diagnosis of thin melanoma is very difficult for the naked eye. This has led to the development of a technique called dermoscopy, which became widely used by dermatologist since the nineties. Dermoscopy is a noninvasive in vivo technique that assists the clinician in melanoma detection in its early stage. Images are acquired using epiluminescence light microscopy, that magnifies lesions and enables examination down to the dermo-epidermal junction. This permits visualization of new morphologic features and in most cases facilitates early diagnosis. However, evaluation of the many morphologic characteristics is often extremely complex and subjective (Rubegni et al., 2002).

In order to make the diagnosis of melanoma based on dermoscopy more objective, three widely used algorithms to analyze melanocytic lesions were proposed in the dermatology literature. All these algorithms consist of identifying a set of features – which are roughly the same for all of them – and classifying the lesion as malignant or benign depending on their absence or presence. While the Menzies method (Menzies et al., 1996) bases its decision on qualitative criteria, the ABCD rule (Stolz et al., 1994; Nachbar et al., 1994) and the 7-point checklist (Argenziano et al., 1998) compute a score. Notice, however, that complete objectivity is impossible to achieve, because of the difficulty in visually characterizing the lesions' features and deciding its presence or absence.

Stolz's ABCD rule specifies a list of visual features associated with malignant melanocytic lesions (asymmetry, border irregularity,

* Corresponding author. Fax: +598 2 711 7435.
E-mail address: pmuse@fing.edu.uy (P. Musé).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.06.015

color irregularity and presence of dermoscopic structures). Non integer coefficients, established based on clinical experience, are associated with each of these features, and the Total Dermoscopic Score (TDS) is computed as the sum of these values. Two thresholds also established by clinical experience are used to classify the lesion as malignant, clinically doubtful (CDL) or benign.

Argenziano's 7-points checklist consists in analyzing the presence of what dermatologist claim to be the seven most important structures that characterize melanoma. These structures consist of color or geometric patterns, that are considered either major or minor criteria. Major criteria structures are weighted by two, while minor criteria structures are weighted by one.

The computerized analysis of dermoscopic images can be an extremely useful tool to measure and detect sets of features from which dermatologists make their diagnosis. It can also be helpful for primary screening campaigns, increasing the possibility of early diagnosis of melanoma. Currently there is no commercial software for massive use in the clinical practice; this evidences that computer aided diagnosis is still an unsolved problem. Our ultimate goal is to develop software for the recognition of early-stage melanoma, using dermatoscopic images. This would enable unsupervised classification of melanocytic lesions, assigning a confidence index for each classification. The result of such classification procedure will separate the "screened" lesions in two groups. The first group corresponds to lesions that were classified with high enough confidence level, while the second one corresponds to those lesions for which the confidence level is low and consequently, requires subsequent inspection by an experienced dermatologist. In this sense, the classification technique is actually a semi-automated method.

In this work we present a computer aided diagnosis method, that results from applying machine learning techniques to attributes inspired on the features used in dermoscopic algorithms. One advantage of the computer over the clinician is that features like those based on asymmetry, border or color irregularity can be quantified with higher precision. Also, instead of deciding on the presence or absence of a given structure, it is possible to quantify its degree of presence, which may turn out to be useful information. The other big advantage of computer aided diagnosis is the ability to infer thresholds and decision boundaries in a more rigorous way using machine learning techniques. On the other hand, the design of computer vision algorithms that perform well in detecting dermoscopic structures, or patterns like those involved in the 7 points checklist, is an extremely difficult task. The reason for this will become evident further, once we have described the aspect of these patterns.

A shorter version of this work was presented in (Capdehourat et al., 2009). In the present paper we show new experiments obtained on a larger database of dermoscopic images. We discuss performance evaluation issues and we choose a different strategy, that we consider to be more accurate than the one adopted in that previous work. We present ongoing work on detection of dermoscopic structures and we propose new guidelines to expand the computer aided diagnosis system reported in (Capdehourat et al., 2009), that include detection of dermoscopic structures. The paper is organized as follows. In Section 2 we present a brief description of dermoscopic algorithms, that suggest the kind of features that have to be extracted from dermoscopic images, and show the limitations of these algorithms. In Section 3 we discuss previous work on computer aided diagnosis of melanoma. Our approach to melanocytic lesions classification is described in Section 4. In Section 5 we present the procedure used to build up a labeled database of dermoscopic images, and we describe its composition. Performance evaluation and results are presented in Section 6. In Section 7 we discuss detection of the dermoscopic structures described in Section 4. Conclusions and future work focused on combining detection of dermoscopic patterns with the classification framework are discussed in Section 8.

2. Description and evaluation of dermoscopy algorithms

2.1. The ABCD rule and the 7 point checklist

The ABCD rule of dermoscopy (Stolz et al., 1994) is based on a scoring system for melanocytic neoplasms that differentiates them into benign, CDL and malignant categories. This is accomplished by computing a TDS according to Table 1. The lesion is considered to be benign if its TDS is lower than 4.75, and malignant if its TDS is larger than 5.45. TDS in between these values correspond to CDL. Features are defined as follows:

- *Asymmetry*: The lesion is bisected by its two principal axes. Symmetry takes into account the contour, colors, and structures within the lesion.
- *Border*: The lesion is divided into octants, by its principal axes and two supplementary axes. Next, one counts the number of segments that have an abrupt border cutoff.
- *Color*: One point per each color listed in Table 1.
- *Dermoscopic structures*: One point per each structure listed in Table 1. Typical examples of these structures can be seen among the patterns marked in Fig. 1.

The 7 points checklist (Argenziano et al., 1998) is another variation on the theme of pattern analysis, with fewer criteria to identify and analyze, and a point system. The dermoscopic patterns are divided into major and minor criteria. Major criteria receive 2 points each and minor criteria receive 1 point (see Table 2). The lesion is considered to be malignant if its total score is larger or equal than 3; otherwise, the lesion is classified as benign. The structures are defined next. Typical examples are shown in Fig. 1.

- *Atypical pigment network*: Black, brown, or gray thickened and irregular line segments anywhere in a lesion.
- *Blue-whitish veil*: Irregular, confluent, gray-blue to whitish-blue diffuse pigmentation that can be associated with pigment network alterations, dots/globules, or streaks.
- *Atypical vascular pattern*: Linear- irregular and/or dotted red vessels not seen in regression areas.
- *Irregular Streaks*: Pseudopods or radial streaming irregularly arranged at the periphery of lesion.

Table 1
ABCD rule. Points in the third column are multiplied by its corresponding weight factor. The TDS is computed as the sum of the sub-scores of each feature. Lesion classification: benign (*TDS* < 4.75), CDL (4.75 ⩽ *TDS* ⩽ 5.45), malignant (*TDS* > 5.45).

Feature	Description	Points	Weight factor	Sub-score range
Asymmetry	One point per asymmetry w.r.t. each axis	0–2	1.3	0–2.6
Border	Eight segments, one point for abrupt pigment cutoff	0–8	0.1	0–0.8
Color	One point per color: white, red, black light brown, dark brown, blue-gray	1–6	0.5	0.5–3
Dermoscopic structures	One point per structure: pigment network, structureless area, dots, globules, branched streaks	1–5	0.5	0.5–2.5
Total score range				1.0–8.9

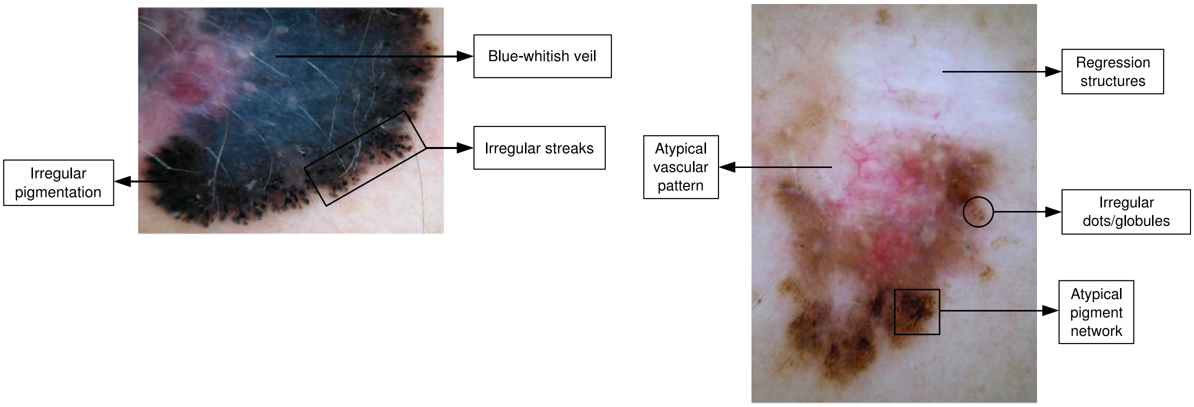

Fig. 1. Dermoscopic structures. Typical examples of the patterns used in dermoscopic algorithms.

Table 2
7 Points checklist. The total score is computed by weighting each structure by its corresponding score, and summing them up. Lesion classification: benign (total score < 3), malignant (total score $\geqslant 3$).

Major criteria	Score	Minor criteria	Score
Atypical pigment network	2	Irregular streaks	1
Blue-whitish veil	2	Irregular pigmentation	1
Atypical vascular pattern	2	Irregular dots/globules	1
		Regression structures	1

- *Irregular pigmentation*: Black, brown or gray featureless areas with irregular shape and/or distribution.
- *Irregular dots/globules*: Black, brown, or gray round to oval, variously sized structures irregularly distributed in the lesion.
- *Regression structures*: white scarlike areas and/or blue pepper-like areas (gray–blue areas, multiple blue–gray dots).

2.2. Performance comparison and implications for computer aided diagnosis

In the ABCD rule, (Stolz et al., 1994) proposed values for the coefficients involved in the TDS and the decision thresholds. While these values are universally used, the optimal ones may be actually quite different; it has been shown that other values may yield better performance, depending on the population under examination and the clinical training. Indeed, many authors claim that the universally used thresholds may lead to high rates of false diagnoses. An experiment conducted by Lorentzen et al. (1999), Lorentzen et al. (2000) revealed that the use of the ABCD rule did not improve diagnostic accuracy of malignant melanoma with respect to qualitative analysis of lesion patterns, for a set of four experienced users of dermoscopy and five less experienced users. Moreover, they present quantitative arguments that support the claim that the points and weights are not correctly balanced.

Concerning the 7 point checklist, (Johr, 2002) claims that this algorithm outperforms the ABCD rule when used by non expert medical doctors. Dolianitis et al. (2005) conduct performance comparison of Menzies method, the ABCD rule and the 7 point checklist, for non expert medical doctors. These authors conclude that Menzies method, which is a qualitative analysis, outperforms the others, and that the 7 point checklist has higher sensitivity but lower specificity than the ABCD rule. Indeed, it is reasonable to think that the diagnosis process conducted by expert dermatologists – as in many medical diagnosis problems – involves several aspects like many years of clinical experience, epidemiology of the disease in different populations, and many more that can be understood as its "medical intuition". Consequently, such a complex decision process cannot be reduced to a simple set of rules.

In sum, a thorough review of the medical literature reveals the lack of consensus in the performance of dermoscopy algorithms. In any case, it seems clear that the points, weights and scores, as well as the thresholds involved in dermoscopy algorithms are not robust neither independent of the context. Solving this issue is the main motivation of this work. Our approach is then focused in taking advantage on the clinical experience to define the features that have to be measured on dermoscopic images, and to leave the determination of optimal weights and decision boundaries to state of the art machine learning techniques.

3. Review of computerized analysis of dermoscopic images

The first related work in the medical literature of computer aided image analysis in skin lesion diagnosis seems to date back to 1987 (Cascinelli et al., 1987). Its contribution was limited since by that time computer vision and machine learning were both emerging fields; note that, for instance, the first widely used edge detector had recently been proposed by Canny in 1986. One of the first significant contributions from the image processing community was reported by Ganster et al. (2001). In this work, the authors propose a classical machine learning approach for dermatoscopic image classification. The first stage is automatic color-based lesion segmentation. Then, over a hundred features that try to reflect parameters used in medical diagnosis are extracted from the image (shape and color, and gradient distribution in the neighborhood of the lesion boundary). Feature selection was obtained using sequential forward and sequential backward floating selection. Classification experiments, performed with a 24-NN classifier, delivered a sensitivity of 77% with a specificity of 84%.

To our knowledge, up to now the best results in automated melanocytic lesion classification were obtained by Celebi et al. (2007). See this reference for a complete summary of the results obtained by key studies from 2001 onwards, along with their database sizes. At this point a precision should be made regarding performance results. First, there is no freely available database, and consequently each work report results on its own database. Second, in general the class corresponding to malignant lesions has far fewer instances than the benign lesion class; in the particular case of Celebi et al. (2007), reported results were obtained using a synthetic oversampling of the minority class, which, as will be discussed later, tends to overestimate performance.

As in (Ganster et al., 2001), the approach proposed by Celebi et al. (2007) is a classic machine learning methodology. After an Otsu-based image segmentation, a set of global features are computed (area, aspect ratio, asymmetry and compactness). Local color and texture features are computed after dividing the lesion in three regions: inner region, inner border (an inner band delimited by the lesion boundary) and outer border (an outer band delimited by the lesion boundary). Feature selection is performed using ReliefF (Robnik-Šikonja and Kononenko, 2003) and CFS algorithms (Hall, 2000). Finally, the feature vectors are classified into malignant or benign using SVM with model selection (Schlkopf and Smola, 2001). Performance evaluation gave a specificity of 92.34% and a sensitivity of 93.33%.

The computer aided diagnosis based on detection of dermoscopic structures has been explored to a much lesser extent, and by a very few research groups. This is certainly due to the fact that the detection of these structures is a difficult problem in image analysis. In a series of papers by a group of University of Salerno (Betta et al., 2006; Di-Leo et al., 2009; Di-Leo et al., 2010), the authors describe methods to detect atypical network, blue-whitish veil, irregular streaks, irregular pigmentation and regression areas, with specificities ranging from 82% and 93%, and sensitivities ranging from 80% to 90%. They do not address detection of atypical vascular patterns neither of dots/globules. Features are computed based on lesion color segmentation (by combining PCA analysis and histogram partitioning with manual pick selection), and texture extraction based on mathematical morphology and Fourier Transform. Detection of each structure is achieved using logistic model trees. The combination of these classifiers to provide a complete lesion classification framework is not addressed. This is a clear limitation, since this combination may yield better performance than using the one or two score per pattern suggested by the 7 point checklist algorithm.

In the following sections we describe our proposed approach, which is mainly based on color, asymmetry and border properties, and give guidelines for combining this approach with dermoscopic structure detection, in a more complete framework.

4. Dermoscopic images classification: Proposed approach

We follow a typical machine learning methodology. In the first stage, we tackle image processing and image analysis problems, such as image filtering, restoration and automatic segmentation to isolate the lesion's area. The second stage consists of extracting features from the image for further lesion classification into malignant or benign. Features are inspired by the same elements that dermatologists use for lesion diagnosis. Once lesions' features have been extracted, labeled lesions are used to train a meta-classifier obtained using boosting based on decision trees. Classification errors and ROC curves are obtained by means of cross validation. In this section we give details of each of these stages.

4.1. Preprocessing and hair removal

Lesion segmentation in the presence of hair is usually doomed to failure. Thus, previous application of a hair removal filter is unavoidable. Ideally, it would be desirable to eliminate hair previous to image acquisition, but this interferes with the clinical practice. In our case, we remove hair using an automatic hair removal algorithm. This algorithm consists of hair detection and image inpainting. For this purpose, we used a well known hair removal algorithm (Lee et al., 1997). This algorithm identifies the image segments that approximate the structure of the hair, and then the regions that contain these segments are interpolated using the information of the surrounding pixels. As for the inpainting, sophisticated state of the art techniques were also explored (Criminisi et al., 2004), with similar results. A typical result is shown in Fig. 2.

4.2. Segmentation

Segmentation of melanocytic lesions can be an extremely hard problem. Besides the presence of hair, many lesions present diffuse borders, difficult to determine even for dermatologists (see Fig. 3). Several methods of image segmentation were explored, based on edge detection and on region information. In general it is appropriate to combine different features (texture, edges, color) for better results. Methods combining these sources of information were also studied. Among the variational methods, we considered a modified version of Otsu (1979) that uses color norm, Mumford-Shah (Koepfler et al., 1994), Geodesic Active Contours (Caselles et al., 1997) and Geodesic Active Regions (Paragios and Deriche, 2002). We also explored several methods based on the topographic map, using both boundary and color/texture region information (Cao et al., 2005; Cardelino et al., 2006).

Overall, none of the methods outperformed the others. We decided to use the color-based Otsu method as it is simpler and significantly faster. Of course, there are pathological cases in which it fails, and sometimes one of the other methods provides satisfactory results. This suggests that a software for clinical use should propose the choice of a few candidate segmentations to the user in case they differ.

4.3. Feature extraction

Once the lesion boundary has been detected, we extract a total number of 57 features to represent the relevant lesion information for melanoma classification. This set of measurements can be divided into global and local features. The subset of global features consist of 9 shape and border features: aspect ratio, symmetry (with respect to principal and secondary axis), circularity, compacity, normalized perimeter, anisotropy, border abruptness and border roughness. The subset of local features represent color and texture information. Following (Celebi et al., 2007), previous to the extraction of local features, each lesion is decomposed into three sub-regions: interior of the lesion and the inner and outer border (Fig. 4). For each of these regions, the color features consist of mean and variance per channel in RGB and HSV spaces, totalizing 36 color descriptors. The remaining 12 texture features, 4 for each sub-region, are based on weighted averages of the gray level co-ocurrence matrix which measures information of contrast, correlation, heterogeneity and energy for each sub-region. Note that information concerning the presence or absence of several geometric patterns that are relevant to the 7 points checklist is not included as attributes in this list of features. We are currently investigating these detection problems, since we are confident that the capability of detecting this structures will boost our method performance. This issue will be addressed in more detail in Section 7.

4.4. Classification

The goal of this stage is to classify the feature vectors in two classes: malignant and benign. A classification technique that proved very successful in our experiments consists in performing decision trees combination *via* adaptive boosting. Boosting exploits the inherent instability of learning algorithms by combining multiple models, in a way that models complement one another. This is achieved by assigning weights to the training data, and modifying them after each classifier by increasing the weight of misclassified samples, and decreasing these of correctly classified ones. Hence, after each iteration, a new classifier is forced to focus on classifying the hard samples correctly. The algorithm finishes after a user-defined number of T iterations, that generates a set of T classifiers. Then, a weight that increases with its performance is associated

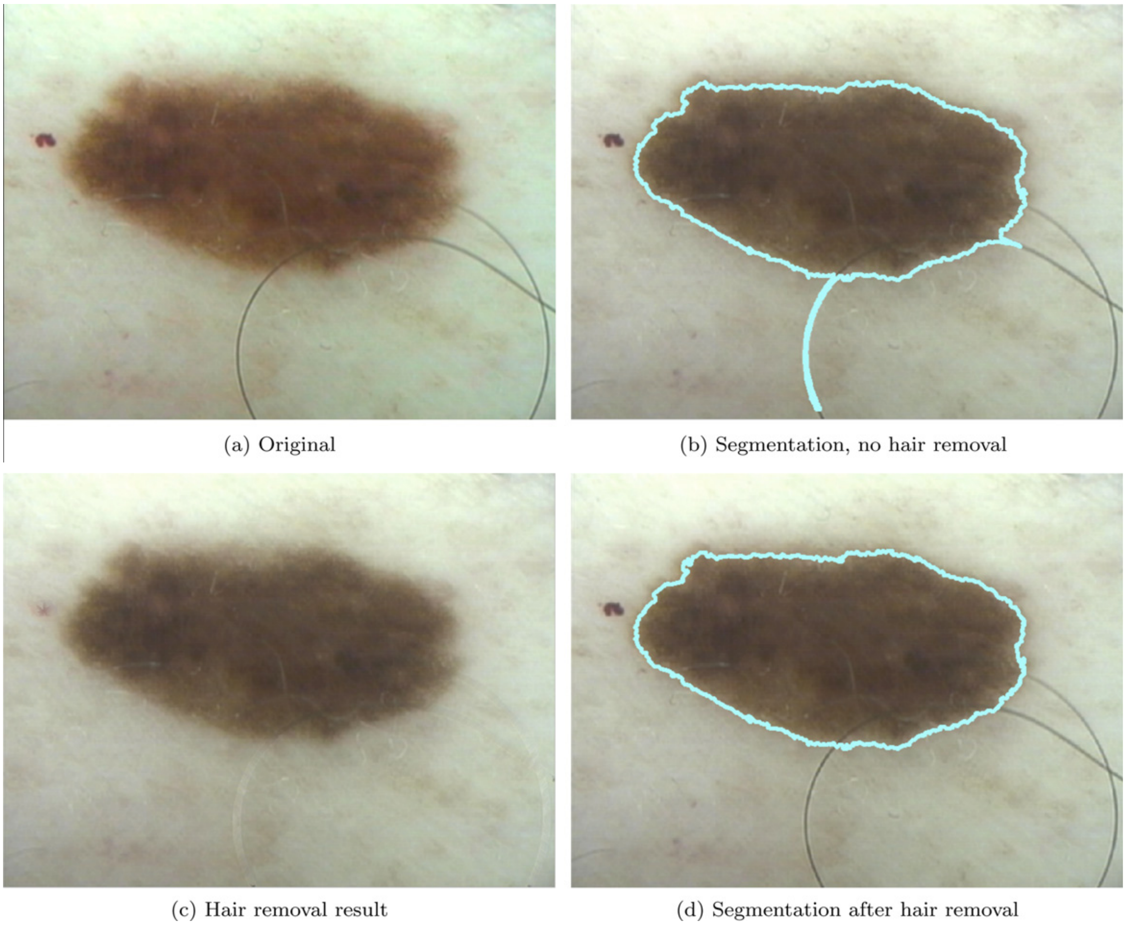
(a) Original
(b) Segmentation, no hair removal
(c) Hair removal result
(d) Segmentation after hair removal

Fig. 2. The importance of hair removal in segmentation. The lesion boundary in (d) is found after applying the hair removal algorithm. Once the lesion has been segmented, the information inside the original lesion (a) is used for subsequent stages.

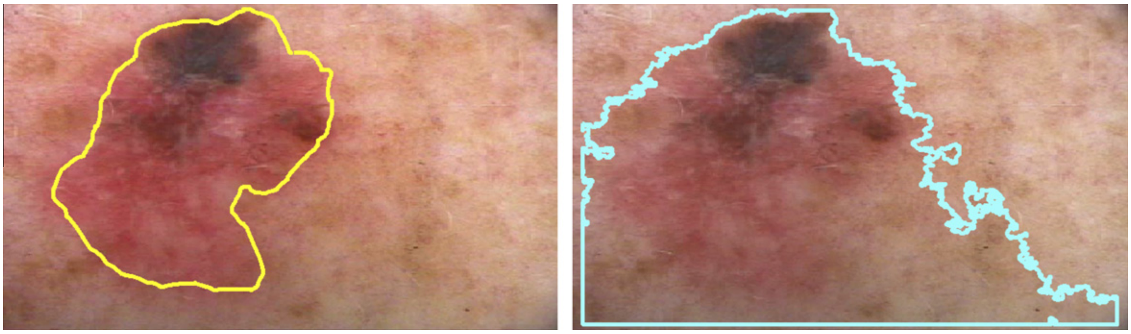

Fig. 3. Example of melanoma showing a very diffuse border. Left: manual segmentation by dermatologist. Right: automatic segmentation with Otsu method based on color norm.

with each of them. Classification of new unlabeled data is performed by a weighted vote of the T classifiers.

The algorithms we considered for the classification framework are C4.5 decision trees (Quinlan, 1993), and AdaBoost.M1 (Freund and Schapire, 1997). Tree classifiers are particularly useful to deal with non-metric data; they exhibit comparable accuracy to widely used classifiers such as neural networks or nearest neighbor classifiers, especially when one does not count on prior information about the appropriate classifier form. Quinlan's C4.5 decision tree has proved to be one of the best performing classification trees. In the C4.5 algorithm the decision tree is grown fully, until leaves have minimum impurity. Nominal values are treated as in Quinlan's ID3 tree, that is by splitting nodes based on optimizing the entropy-based information gain. Continuous variables are treated as in Brieman's Classification And Regression Trees (CART), that is, by sorting the values and choosing the splitting point as the one that optimizes the information gain. Then, the full grown tree is subject to cost-complexity pruning, which has the ability of directly replacing a complex subtree with a leaf.

AdaBoost (ADAptive BOOSTing) is one of the many variations on basic boosting. It allows the designer to continue adding weak classifiers until reaching some target training error. In AdaBoost each

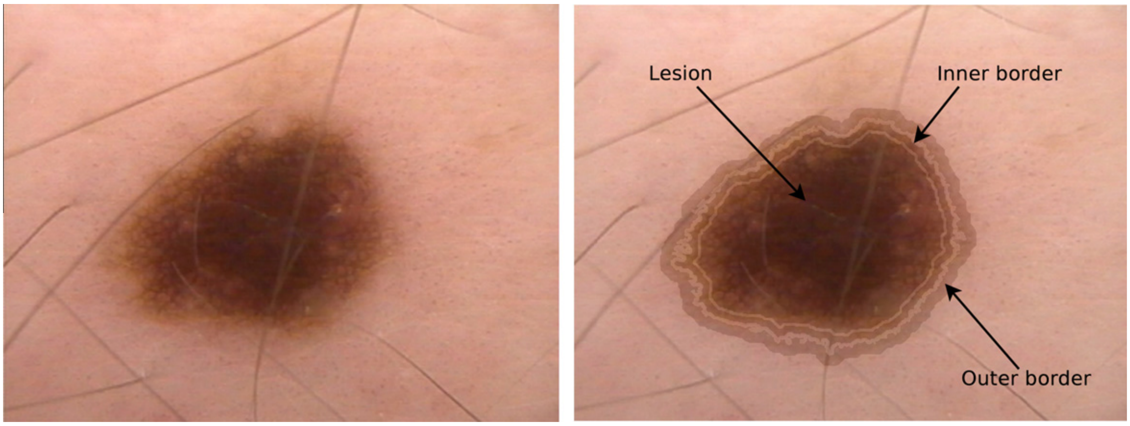

Fig. 4. Definition of the three sub-regions used in local features extraction.

training sample receives a weight which determines its probability of being selected for a training set for an individual component classifier. If a training sample is accurately classified, then its chance of being used again in a subsequent component classifier is reduced; conversely, if the sample is not classified accurately, then its chance of being used again is raised. In this way, AdaBoost concentrates on those samples that are hard to classify. In the initialization, all training samples weights are set to the same value. Then, on each iteration, training samples are drawn at random according to these weights, and train a component classifier (here a C4.5 decision tree) using the selected samples. Next, the weights of the samples that were correctly classified are decreased, and those of misclassified samples are increased. Then a new subset of samples, drawn following this updated distribution, is used to train the next classifier. The process is iterated until a design training error is reached.

5. Database composition

In order to enable performance evaluation for lesion segmentation, features' measurements and melanoma classification, a dermoscopy image database was built. To simplify the work of the dermatologist in labeling the database, a graphical tool for manual segmentation and diagnosis was developed. For each database image, the dermatologist defines the lesion boundary by hand, and digitally fills a diagnostic report for both dermoscopy algorithms (ABCD rule and 7 points checklist). An expert dermatologist processed our complete set of dermoscopic images using this tool, leading to a full labeled database.

The database is composed of 655 images of melanocytic lesions: 544 benign lesions and 111 malignant melanoma. Actually, the original set of dermoscopic images was larger, but some images were discarded for the following reasons: the images do not capture the whole lesion, poor image quality or excessive presence of hair. Among the set of benign melanocytic lesions, 150 correspond to dysplastic nevi, 77 dermal nevi, 36 junctional nevi, 65 compound nevi and 216 unclassified. This composition was based on the existence of dermoscopic and histopathologic studies, which were used as ground truth for the classification procedure. It is important to note that dysplastic melanocytic nevi are the benign lesions that are visually the most alike to malignant melanoma; many of them are clinically doubtful for experienced dermatologists.

6. Performance evaluation and results

Performance evaluation was conducted using 10 times – 10-fold cross-validation. To assess the impact of the learning and classification method, we compared our results with SVM with model selection (preceded by ReliefF feature selection). A RBF kernel was used, and optimal parameters (the weight that controls model complexity and the RBF parameter) were obtained by grid search optimization with 10 fold cross-validation. The same experiments were repeated, replacing automatic segmentation with manual segmentation, performed by a dermatologist. This was carried onto assess the influence of automatic segmentation errors.

In a previous version of this work (Capdehourat et al., 2009), in order to deal with class imbalance in a situation where the size of the minority class was small, we applied on it a widely used synthetic over-sampling technique (SMOTE, by Nitesh V. Chawla et al. (2002)). This enabled us to compare our results with those reported by Celebi et al. (2007), where the same over-sampling technique was used. Note that since the database used by Celebi et al. is very similar to ours in size and composition (476 benign lesions and 88 malignant melanoma), this performance comparison makes sense, but only up to a certain point. The results we obtained in this previous work were better than those reported by Celebi et al. (2007) (who obtain a specificity of 86% for a 95% sensitivity, and AUC of 0.966). In the experiments reported here, with a larger database and still using SMOTE to compensate class imbalance, our AdaBoost/C4.5 approach shows again higher performance (specificity of 89% for a 95% sensitivity, and AUC of 0.977).

More recently, in a joint exploration with Fiori et al. (2010), we have observed that the use of SMOTE tends to overestimate the performance of the classification method. For example, generating five features with independent, identically distributed random values (standard normal distributions for both classes), and classifying with Naive Bayes, values of AUC of almost 0.8 were reached (when 0.5 was expected). For that reason, partly because now the number of melanomas in the database is larger, we decided to conduct performance evaluation by randomly subsampling the majority class. This procedure was repeated in order to use all instances in the majority class.

Fig. 5 shows the overall system performance (ROC curves) using AdaBoost/C4.5 for both automatically and manually segmented databases. The plot on the top shows the results when applying SMOTE to deal with class imbalance. The bottom plot shows the corresponding results when using subsampling of the majority class. Note that the use of SMOTE clearly biases the system performance evaluation. In both cases, results obtained over the manually segmented database are slightly superior. Table 3 shows performance indicators, for the manually and automatically segmented databases, using Adaboost/C4.5 and SVM, and both class balancing strategies. While the SVM approach using manually or automatically segmented images yielded essentially the same performance, the performance of Adaboost/C4.5 classification of man-

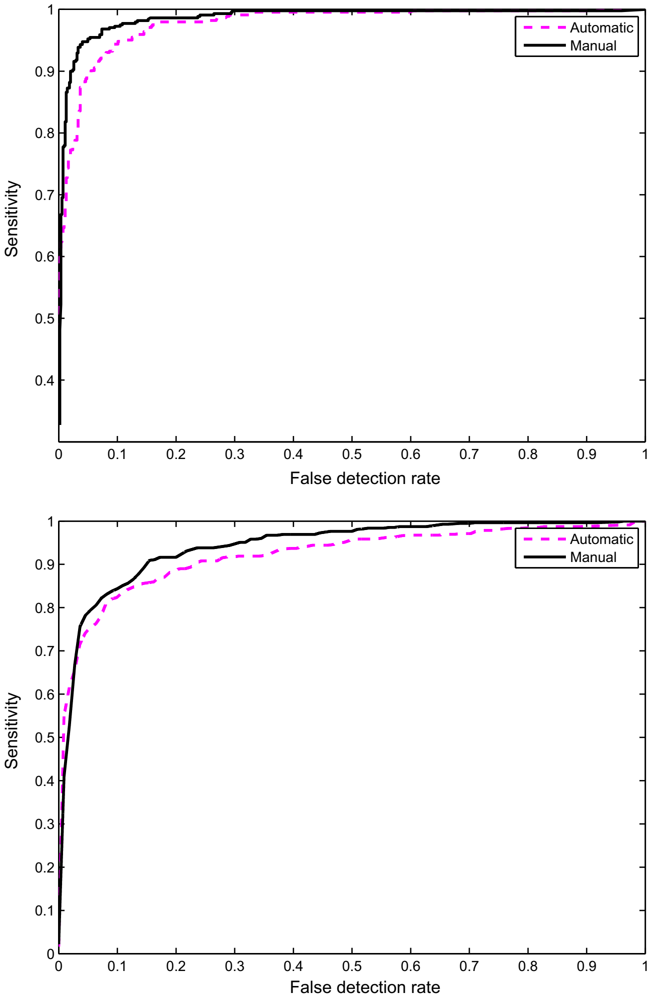

Fig. 5. ROC curves for automatic and manual lesion segmentation. Top: oversampling of the minority class using SMOTE. Bottom: random subsampling of the majority class.

Table 3
Performance indicators for two different classifiers, and two class balancing strategies.

Segmentation	Classifier	SMOTE		Spread subsampling	
		Specificity for 90% sensitivity (%)	AUC	Specificity for 90% sensitivity (%)	AUC
Automatic	AdaBoost – C4.5	94.5	0.977	77.0	0.921
	SVM	91.1	0.953	74.7	0.890
Manual	AdaBoost – C4.5	97.8	0.986	85.0	0.937
	SVM	89.5	0.948	75.0	0.889

ually segmented images was higher than for the automatically segmented ones. In agreement with our previous experiments (Capdehourat et al., 2009), the classification results using Adaboost/C4.5 are better than the ones obtained with SVM, for all cases. To end up with this section, Table 4 shows performance indicators for dermoscopy algorithms reported by Dolianitis et al. (2005). Comparison with the ROC curves in Fig. 5 reveals that our method outperforms all of them. Again, note that this comparison should not be taken *stricto sensu*, because of multiple factors (different databases, etc.).

Table 4
Performance indicators for dermoscopy algorithms, reported by Dolianitis et al. (2005).

Dermoscopy algorithm	Sensitivity	Specificity
ABCD rule	77.5	80.5
7 points checklist	81.4	73.0
Menzies	84.6	77.7

7. A preliminary study on detection of dermoscopic structures

In this section we address the detection of three dermoscopic structures used in the 7 points checklist. A simple blue-whitish veil detector, proposed by Celebi et al. (2008), was implemented and tested. For the detection of atypical pigment network, a simple detector is proposed. Finally, we propose a new detector of irregular pigmentation, that combines three detectors reported in the literature, and provides a binary answer regarding the absence or presence of such structure. The goal of this study is to evaluate the state of the art in this area, and to gain insight with pattern detection in order to develop new classifiers based on these dermoscopic structure detectors. The final objective is to combine these classifiers based on dermoscopic structures, with the classifier proposed in the previous sections, which is based on general shape, color and texture features.

7.1. Blue-whitish veil

A blue-whitish veil detector was implemented, based on Celebi et al. (2008). In this work the classification of each pixel as veil or non veil is obtained using a decision tree over two color features extracted from the image. This algorithm was tested on a selected subset of 39 images from our database, with encouraging results. On this subset, this detector did not miss any lesion presenting blue-whitish veil. Figs. 6 and 7 show examples of successful and false blue-white veil detections, respectively.

7.2. Atypical pigmented network

Di-Leo et al. (2010) proposed a strategy for atypical pigmented network detection. This strategy is based on texture features, extracted from the lesion using mathematical morphology and Fourier Transform. From these features, a network image like the one shown in Fig. 8 is produced. Presence or absence of atypical pigment network is decided by means of a logistic model tree fed by statistics extracted from the network image. We implemented another algorithm to generate the network image, based also on mathematical morphology. The method consists in finding segments in the image, in a similar way to the first stage of the hair removal algorithm described in Section 4. A network image example obtained with this method is shown in Fig. 9. Not surprisingly, results are satisfactory because this algorithm tries to identify image segments that approximate the structure of the network.

7.3. Irregular pigmentation

We consider an approach to atypical pigmentation detection based on three detectors of asymmetric blotches (or structureless areas), adapted from (Stoecker et al., 2005; Pellacani et al., 2003). Several modifications were introduced. While in their work the authors calculate geometric features of the blotches found to characterize lesion malignancy, our aim is slightly more ambitious as

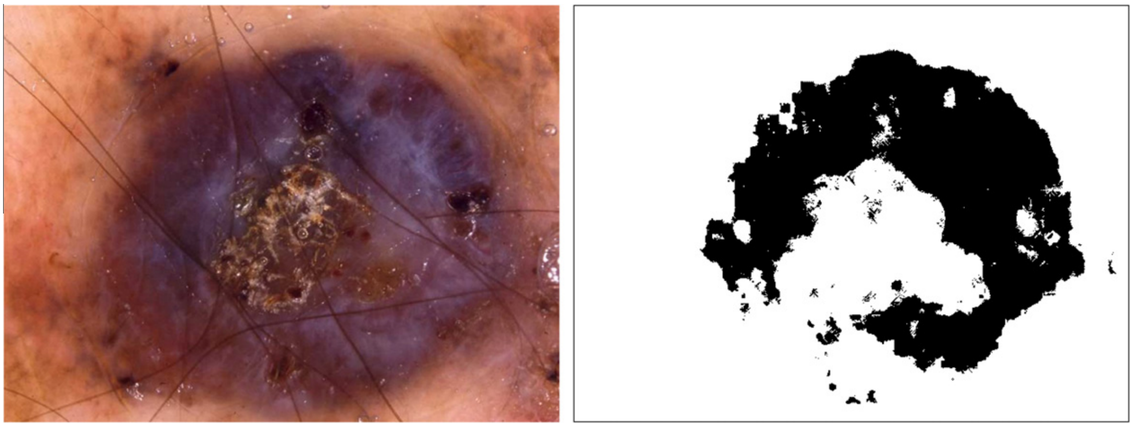

Fig. 6. Left: original image with blue-whitish veil. Right: detected blue–white veil area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

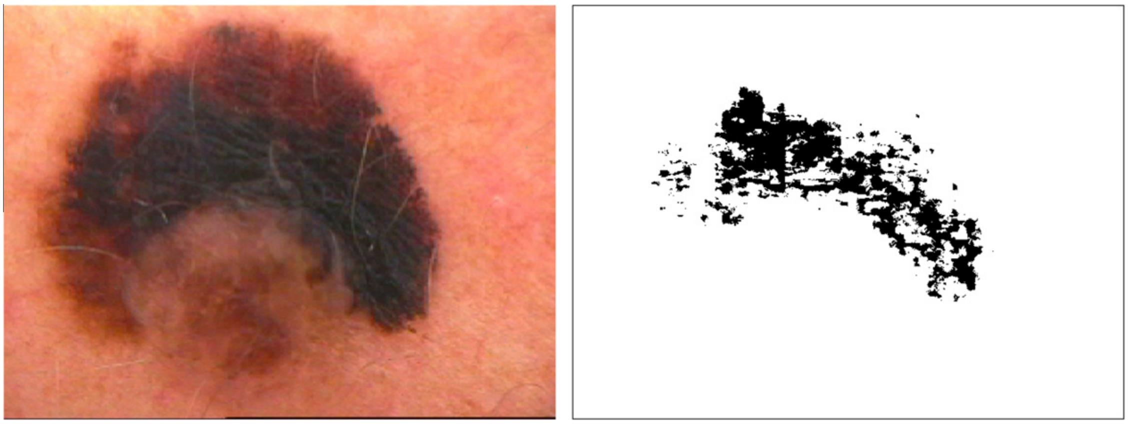

Fig. 7. Left: original image that does not present blue-whitish veil. Right: detected blue–white veil area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

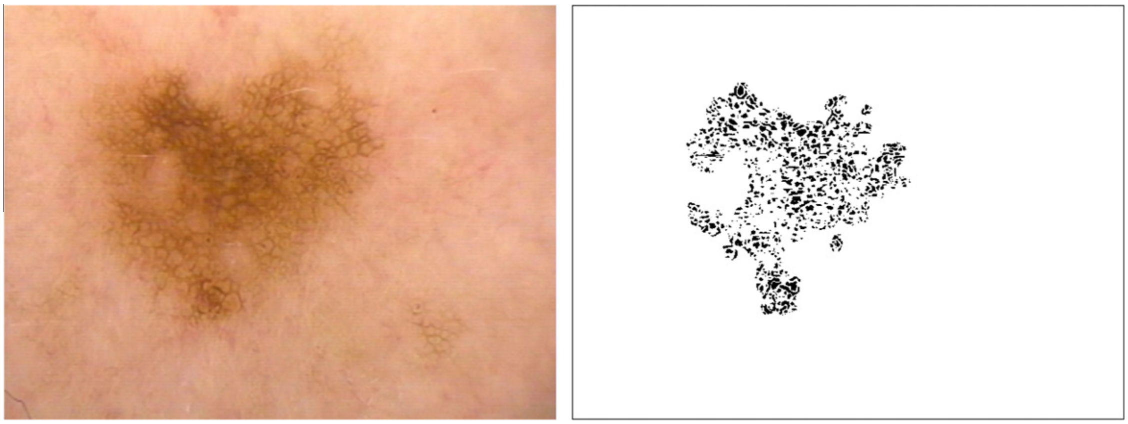

Fig. 8. Left: original image. Right: detected atypical pigmented network area.

we focus on determining the presence or absence of irregular pigmentation.

The binary outputs of the three classifiers (presence or absence of irregular pigmentation) are combined for better results. The two algorithms based on (Stoecker et al., 2005) perform structure detection by simply thresholding the red and green channels. In one of these algorithms, each pixel within the lesion is compared to the fixed threshold. In the other one, the difference between the color of lesion pixels and the outer skin color is considered. The third detector is based on (Pellacani et al., 2003), which only uses gray level images, and detects dark areas by simply thresholding the gray level; we defined the threshold value by choosing the value that maximized the classification performance (presence or absence of irregular pigmentation) when ran on the entire labeled

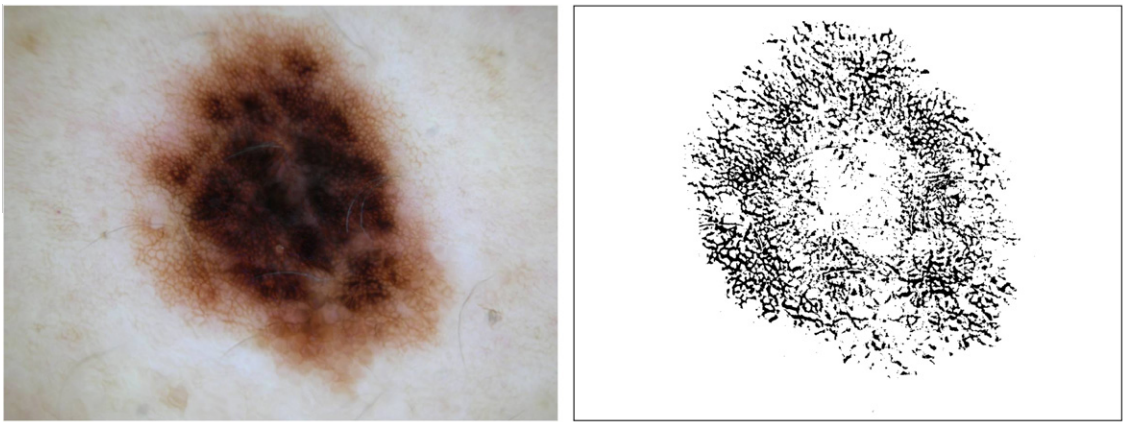

Fig. 9. Left: Original image. Right: network image obtained with the proposed mathematical morphology-based algorithm.

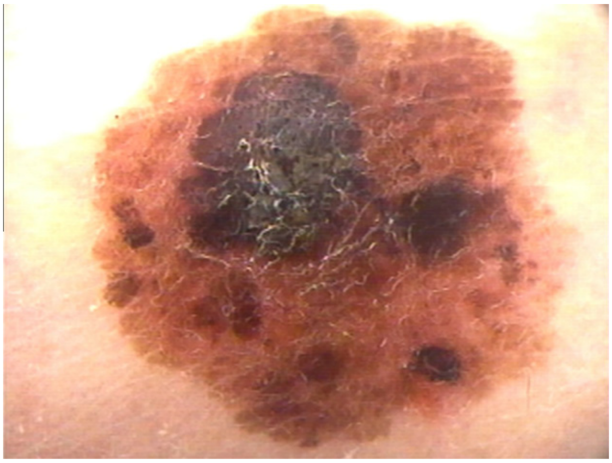

Fig. 10. Example of lesion presenting irregular pigmentation, that is correctly detected.

database. In all three binary classifiers, pigmentation is typified as atypical if any of the two following conditions on the largest detected connected component hold: (i) its area exceeds half the lesion size; (ii) its area falls within the range of 10% to 50% of the lesion size, and the distance between its barycenter and the lesion's barycenter is larger than 20% of the lesion diameter. Using this criteria, each of the classifiers provide a yes/no answer to the presence of irregular pigmentation. Then, the three classifiers are combined using Adaboost. Performance evaluation of irregular pigmentation detection, on the entire database, gave an AUC of 0.66.

Fig. 10 shows a lesion that is well suited for the detection algorithm described above. The main difficulties in characterizing atypicality, that are not taken into account by the three classifiers described above, are the existence of irregular pigmentation which is not dark enough (Fig. 11, top), and on the contrary, quite dark lesions with no irregular pigmentation (Fig. 11, bottom). We are currently working on these problems, and investigating new algorithms, such as (Madasu and Lovell, 2009; Heckbert, 1982; Di-Leo et al., 2010).

8. Conclusions and future work

In this work we presented a machine learning approach to classify melanocytic lesions from dermatoscopic images. Feature extraction is inspired from popular dermoscopic algorithms described in the medical literature. The use of machine learning techniques allows to overcome major limitations of dermoscopy algorithms, namely ad hoc choices of weights and thresholds involved in the decision process. The learning and classification stages are performed using AdaBoost with C4.5 decision trees. Using automatically segmented images, we obtained a specificity of 77% for a sensitivity of 90%, and an AUC of 0.921. It seems also, from the comparison of the results obtained from manually segmented lesions (specificity of 85% for a sensitivity of 90%, AUC of 0.937), that errors in automatic segmentation have an important impact and should be reduced. As pointed out earlier, this is a hard problem since many melanocytic lesions show highly diffuse contours. Note, however, that nothing prevents us to manually seg-

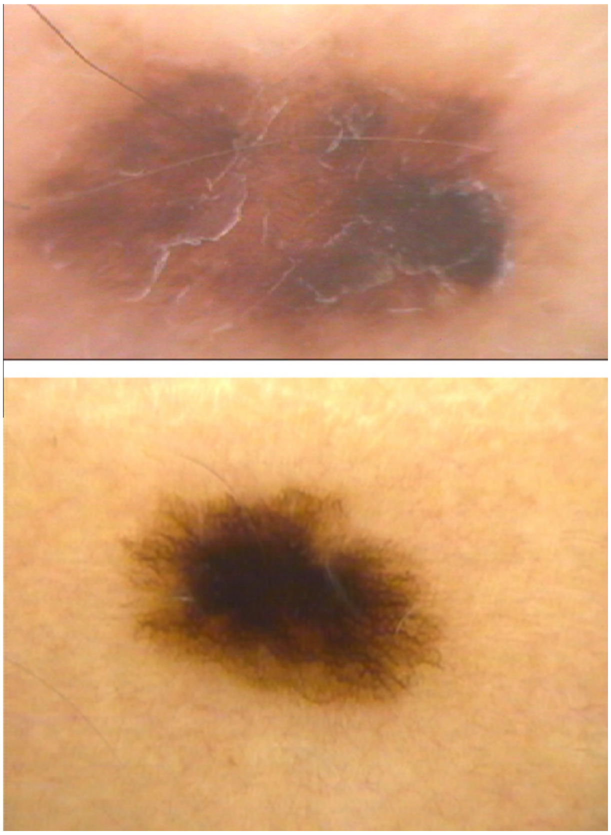

Fig. 11. Two lesions that are not well suited for detection with the proposed algorithm. Top: irregular pigmentation with low contrast. Bottom: dark lesion that does not present irregular pigmentation.

ment the training database, and to propose to the user, for each new lesion, the choice of candidate segmentations.

Results of the proposed approach are promising and seem to be superior than those reported in the literature. However, performance evaluation is delicate because all reported results were obtained using different databases and different validation strategies. At this point, construction of a large database of dermatoscopic images that could be used as reference testbed appears to be a fundamental issue.

Concerning our algorithm, to further improve its performance, methods to detect a larger number of geometry or texture based structures, similar to those used in the 7 points checklist, should be developed. Because of their strong discriminative power, we are confident that the inclusion of these patterns information in the classification framework will boost the performance. The detection of dermoscopic patterns is ongoing research; a preliminary study was presented here. The next step will be to design a classifier based on the detection of these structures.

Note that, contrarily to the classifier based on asymmetry, color/texture and border irregularities proposed in the previous sections, the information related to dermoscopic patterns cannot be directly used for melanoma detection. Indeed, the presence of all the patterns is not systematic to every melanoma. This suggest that a separate classifier should be designed for the dermatoscopic structures-based approach, and both results have to be presented to the clinician for further diagnosis. Hopefully this combination strategy will be implemented in future versions. To our knowledge, such a combined framework for automatic melanocytic lesion classification has not been proposed yet.

Another interesting related line of research is the characterization of the discriminative power of the considered features. This can be obtained by means of automatic feature selection strategies like the ones that were mentioned here. A rigorous study of this topic, complemented with the comparison of the weights assigned to visual features in the ABCD and other clinical diagnosis rules, may yield useful recommendations to dermatologist for their medical practice. As evidenced by an exhaustive exploration of the medical literature, the choice of weights and thresholds involved in the dermoscopy algorithms is far from being a solved issue.

Acknowledgments

This work was partially supported by CSIC – Universidad de la República, Comisión Honoraria de Lucha Contra el Cancer, Uruguay, and the Uruguayan Agency for Research and Innovation (ANII) under grant PR-POS-2008-003.

References

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions, comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch. Dermatol. 134, 1563–1570.

Betta, G., Di-Leo, G., Fabbrocini, G., Paolillo, A., Sommella, P., 2006. Dermoscopic image-analysis system: Estimation of atypical pigment network and atypical vascular pattern. MEMEA '06: Proceedings of the IEEE International Workshop on Medical Measurement and Applications, 2006. IEEE Computer Society, Washington, DC, USA, pp. 63–67.

Cao, F., Musé, P., Sur, F., 2005. Extracting meaningful curves from images. J. Math. Imaging Vision 22 (2-3), 159–181.

Capdehourat, G., Corez, A., Bazzano, A., Musé, P., 2009. Pigmented skin lesions classification using dermatoscopic images. CIARP '09: Proceedings of the 14th Iberoamerican Conference on Pattern Recognition. Springer-Verlag, Berlin, Heidelberg, pp. 537–544.

Cardelino, J., Randall, G., Bertalmio, M., Caselles, V., 2006. Region based segmentation using the tree of shapes. In: Proc. IEEE Internat. Conf. on Image Process.

Cascinelli, N., Ferrario, M., Tonelli, T., Leo, E., 1987. A possible new tool for clinical diagnosis of melanoma: The computer. J. Amer. Acad. Dermatol. 16 (2), 361–367.

Caselles, V., Kimmel, R., Sapiro, G., 1997. Geodesic active contours. Internat. J. Comput. Vision 22 (1), 61–79.

Celebi, M.E., Iyatomi, H., Stoeckler, W.V., Moss, R.H., Rabinovitz, H.S., Argenziano, G., Soyer, H.P., 2008. Automatic detection of blue–white veil and related structures in dermoscopy images. Comput. Med. Imaging Graph.: Official J. Comput. Med. Imaging Soc. 32, 670–677.

Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker, W.V., Moss, R.H., 2007. A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. 31 (6), 362–373.

Criminisi, A., PTrez, P., Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 13, 1200–1212.

Di-Leo, G., Fabbrocini, G., Paolillo, A., Rescigno, O., Sommella, P., 2009. Estimation of chromatic parameters for automatic diagnosis of skin lesions. MASAUM J. Comput. 1 (3), 369–376.

Di-Leo, G., Paolillo, A., Sommella, P., Fabbrocini, G., 2010. Automatic diagnosis of melanoma: A software system based on the 7-point check-list. HICSS. IEEE Computer Society, pp. 1–10.

Dolianitis, C., Kelly, J., Wolfe, R., Simpson, P., 2005. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. Arch. Dermatol. 141 (8), 1008–1014.

Fiori, M., Musé, P., Aguirre, S., Sapiro, G., 2010. Automatic colon polyp flagging via geometric and texture features. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10 (August 2010), pp. 3170–3173.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

Ganster, H., Pinz, A., Rhrer, R., Wildling, E., Binder, M., Kittler, H., 2001. Automated melanoma recognition. IEEE Trans. Med. Imaging 20, 233–239.

Hall, M.A., 2000. Correlation-based feature selection for discrete and numeric class machine learning. In: ICML '00: Proc. Seventh Internat. Conf. on Mach. Learning. San Francisco, CA, USA, pp. 359–366.

Heckbert, P., 1982. Color image quantization for frame buffer display. SIGGRAPH Comput. Graph. 16 (3), 297–307.

Johr, R.H., 2002. Dermoscopy: Alternative melanocytic algorithms – The abcd rule of dermatoscopy, menzies scoring method, and 7-point checklist. Clinics Dermatol. 20 (3), 240–247.

Koepfler, G., Lopez, C., Morel, J.M., 1994. A multiscale algorithm for image segmentation by variational method. SIAM J. Numer. Anal. 31 (1), 282–299.

Lee, T., Ng, V., Gallagher, R., Coldman, A., 1997. Dullrazor: A software approach to hair removal from images. Comput. Biol. Med. 27 (11), 533–543.

Lorentzen, H., Weismann, K., Kenet, R., Secher, L., Larsen, F., 2000. Comparison of dermatoscopic ABCD rule and risk stratification in the diagnosis of malignant melanoma. Acta Derm. Venereol. 80 (2), 122–126.

Lorentzen, H., Weismann, K., Secher, L., Petersen, C., Larsen, F., 1999. The dermatoscopic ABCD rule does not improve diagnostic accuracy of malignant melanoma. Acta Derm. Venereol. 79, 469–472.

Madasu, V.K., Lovell, B.C., 2009. Blotch detection in pigmented skin lesions using fuzzy co-clustering and texture segmentation. Digital Image Comput.: Tech. Appl. 0, 25–31.

Menzies, S., Ingvar, C., Crotty, K., McCarthy, W., 1996. Frequency and morphologic characteristics of invasive melanoma lacking specific surface microscopic features. Arch. Dermatol. 132, 1178–1182.

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., Plewig, G., 1994. The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. J. Amer. Acad. Dermatol. 30 (4), 551–559.

Nitesh V. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. Smote: Synthetic minority over-sampling technique. J. Artif. Intell. Res. (JAIR) 16, 321–357.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybernet. 9 (1), 62–66.

Paragios, N., Deriche, R., 2002. Geodesic active regions: A new framework to deal with frame partition problems in computer vision. J. Visual Commun. Image Representation 13, 249–268.

Pellacani, G., Grana, C., Cucchiara, R., Seidenari, S., 2003. Automated extraction and description of dark areas in surface microscopy melanocytic lesion images. Dermatology 208, 21–26.

Quinlan, R.J., 1993. C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning). Morgan Kaufmann.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. 53 (1–2), 23–69.

Rubegni, P., Burroni, M., Dell'eva, G., Andreassi, L., 2002. Digital dermoscopy analysis for automated diagnosis of pigmented skin lesion. Clinics Dermatol. 20 (3), 309–312.

Schlkopf, B., Smola, A.J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press.

Soyer, H., Argenizano, G., Hofmann-Wellenhof, R., Johr, R. (Eds.), 2007. Color Atlas of Melanocytic Lesions of the Skin. Springer.

Stoecker, W., Gupta, K., Stanley, R., Moss, R., Shrestha, B., 2005. Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color. Skin Res. Technol. 11, 179–184.

Stolz, W., Riemann, A., Cognetta, A., Pillet, L., 1994. Abcd rule of dermatoscopy: A new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. 4, 521–527.