Two Systems for the Detection of Melanomas in Dermoscopy Images using Texture and Color Features

Catarina Barata¹ Margarida Ruela¹ Mariana Francisco¹ Teresa Mendonça² Jorge S. Marques¹

Abstract—Melanoma is one of the deadliest forms of cancer hence, a great effort has been put in the development of diagnosis methods for this disease. This paper addresses two different systems for the detection of melanomas in dermoscopy images. The first system uses global methods to classify skin lesions, while the second one uses local features and the Bag of Features classifier. This paper aims at determining which is the best system for skin lesion classification. The other objective is to compare the role of color and texture features in lesion classification and which set of features is more discriminative. It is concluded that color features outperform texture features when used alone and that both methods achieve very good results: Sensitivity = 96% and Specificity = 80% for global methods against Sensitivity = 100%, Specificity = 75% for local methods. The classification results were obtained on a dataset of 176 dermoscopy images, from Hospital Pedro Hispano - Matosinhos.

Index Terms—Melanoma, Dermoscopy, Computer Aided Diagnosis, Bag of Features, Color, Texture.

I. INTRODUCTION

THE American Cancer Society estimates that more than 76000 new cases of melanoma will be diagnosed in 2012. This type of skin cancer is the less common, accounting for less than 5% of all skin cancer cases. However, it is by far the most aggressive since it is more likely to metastasize than other skin tumors. This characteristic makes melanoma the deadliest form of skin cancer (it is estimated that more than 75% of the deaths related with skin cancer in 2012 will be from melanoma) [1]. Melanoma incidence rates have been increasing significantly in the last decades, which makes this one of the cancers that has been receiving attention both from the public health field, with medical prevention campaigns, and from the cancer research field.

The ultimate goal for physicians is to diagnose melanoma in its earliest stage, since it is less probable that it has already metastasize and thus greatly increasing the probability of survival. Nowadays, a technique used by dermatologists to diagnose skin lesions and, consequently, to detect melanomas is Dermoscopy. This is a non-invasive procedure used for the in-vivo observation of skin lesions. The physician places a gel on the skin lesion and inspects it with a magnification instrument (dermatoscope, stereomicroscope or a digital imaging system), which amplifies the lesion 6-100x, depending on the instrument used [2]. This magnification allows the recognition of several surface and subsurface structures, which are not visible by the naked eye, and that can be used to diagnose a skin lesion using one of the several medical diagnostic algorithms: e.g., ABCD rule [3], Menzies method [4] and 7-point checklist [5]. These three algorithms have a common first step: the identification of the inspected lesion regarding its origin as melanocytic or non melanocytic - melanoma is a melanocytic lesion, since it is a malignant proliferation of melanocytes. In their second step, they distinguish between benign melanocytic lesions and melanoma using different approaches. ABCD rule assigns a score to a lesion. This score is a combination of the scores for four different features: (A)symmetry, (B)order, number of different (C)olors and atypical (D)ifferential structures (atypical pigment network, dots, streaks, etc). A score higher than 5.45 diagnoses a lesion as melanoma [3]. Menzies method identifies two types of dermoscopic features: negative (symmetrical pattern, single color) and positive (blue-white veil, atypical dots and network, etc). The presence of positive features signs a melanoma [4]. Finally, the 7-point checklist also scores a lesion. However, this algorithm only inspects the lesion for the presence of atypical differential structures. The scores for different dermoscopic structures can be 2 or 1 and are added up to give the final score. A score higher than 3 indicates melanoma [5]. Fig. 1 illustrates a melanoma and some of the common dermoscopic features considered by
Dermoscopy can actually increase the sensitivity of melanoma detection by 10-27% [6]. However, it has been demonstrated that Dermoscopy only increases the diagnostic performance if the dermatologists have received formal training [7]. Moreover, even with the use of the described diagnostic algorithms which aim to make the diagnosis more reliable and reproducible, the diagnosis of a skin lesion by a dermatologist is still a subjective one since it depends on human vision and on the clinical experience. Computerized dermoscopy image analysis systems can be used to tackle this problem. In these systems, a computer is used as a diagnostic tool and can also be used to follow-up suspicious skin lesions.

This paper proposes two systems for the automatic classification of melanocytic skin lesions. The first system uses global methods to classify skin lesions. This approach evolves in three sequential steps. First, the lesion is segmented using an automatic segmentation method. Then, a set of features inspired in the ABCD rule (color and texture features) is extracted and used to train a classifier to perform a binary classification as melanoma or benign. The second system uses local features, following a recent trend in image analysis and recognition. The paper is organized as follows. Section II gives an overview of several systems proposed in literature. Section III provides an overall description of global and local strategies. Sections IV and V describe the proposed systems and the type of features extracted. The results obtained and their discussion are performed in Section VI. Section VII concludes the paper with some remarks on future work.

II. RELATED WORKS

In the last decade, several systems for melanoma detection have been proposed. Some systems try to mimic the performance of dermatologists by detecting and extracting several dermoscopic structures, such as pigment network [8], [9], [10], irregular streaks [11], granularities [12], regression structures [11], blue-white veil [13] and blotches [14]. These structures can then be used to score a lesion in a similar way to the one adopted by dermatologists. A CAD system which mimics the performance of the 7-point checklist method can be found on literature [15].

However, the vast majority of melanoma detection systems found on literature follow a pattern recognition approach [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. Most of these works use Global Methods to classify skin lesions [16], [17], [18], [19], [20], [22], [24]. The described systems usually consist of three/four consecutive steps: lesion segmentation, feature extraction, feature selection and lesion classification using a trained classifier. These systems are usually inspired in the ABCD rule [3] and the extracted features aim to reproduce each one of the accounted scores. The most common features used in these studies include shape features (e.g., compactness, aspect ratio and maximum diameter), which represent both asymmetry and border, color features in several color spaces (e.g., mean and standard deviation) and texture features (e.g., gray level co-occurrence matrix) [22]. Interesting results have been obtained by different authors with different sets of features and classifiers. Ganster et al. [16] extracted shape and color features and used a k-Nearest Neighbor classifier to distinguish between melanoma and benign nevi. They used a large dataset with more than 5300 dermoscopy images and achieved a Sensitivity (SE) of 87% and a Specificity (SP) of 92%. Rubegni et al. [18] also used shape and color features and added texture features. They achieved a SE = 96% and a SP = 93% in a dataset with 217 melanomas and 588 images, using an Artificial Neural Network (ANN) classifier. Celebi et al. [22] proposed a system based on Support Vector Machines (SVM) classifier to identify melanomas. They used a large feature vector which contained shape, color and texture features and achieved a SE = 93% and SP = 92%. The system proposed by Iyatomi et al. [24] is an internet-based melanoma screening system that uses a feature vector of over 400 features, divided in shape, texture and color features. Their dataset contained more than 1200 dermoscopy images and they achieved a SE and SP both equal to 86%. Local Methods have also been recently proposed to classify skin lesions. Situ et al. [25] describe an algorithm for lesion classification which uses the Bag-of-Features approach. They represent each image as a set of several patches sampled from a 16×16 regular grid placed on the lesion. To described each patch they use Wavelets and Gabor-like filters, leading to a total of 23 features. Two different classifiers were compared: Naive Bayes classifier and SVM; and their best performance is 82% on a dataset of 100 dermoscopy images, 30 of which were melanomas.

The works described above consider that the three commonly used types of features (shape, color and texture) are equally relevant for the automatic detection of melanomas. Other works use a single class of features as a descriptor for melanoma detection [26], [21], [27], [23], [28]. Color distribution in the RGB color space (mean RGB distance, variance and maximum distance) were used to distinguish between melanomas, atypical nevi and benign nevi by Seidenari et al. [21]. They achieved an accuracy of 86% and observed significant difference in pigment distribution between the three studied populations. Relative color histograms have also been used to distinguish between melanomas and benign or atypical nevi [26], [23], both with SE greater than 80%. The Gray level Co-occurrence matrix and other texture descriptors, such as Laws energy masks and Gabor-like filters were applied to classify skin lesions using texture information only [28], [27]. The most promising results were achieved by Sheha et al. [28], who obtained an accuracy of 92% on a dataset of 102 images (51 melanomas). Despite being very promising and demonstrative that both types of features are good discriminators, these cannot be compared since different datasets and experimental conditions are employed. Therefore, it is not possible to determine the relative importance of each feature for skin lesion classification. To the best of our knowledge no work can be found on literature which compares the relevance of each type of features. Furthermore, most works describe the lesion using global features although it has been shown that local features perform better in several image analysis problems (e.g., object recognition and image retrieval) [29].
This paper addresses two distinct objectives. First, it aims at determining which is the best strategy for skin lesion classification: global or local? Second, it aims to compare the performance of two classes of features (color and texture) in melanoma identification and determine which feature is the best discriminator in this specific problem.

III. SYSTEM OVERVIEW

The analysis of dermoscopy images is an object recognition problem under very specific conditions. Several approaches have been proposed to deal with this kind of problems. Some of them characterize the object by a set of global features (e.g., color moments and gradient histograms) and use these features to discriminate the object from the background. However, these techniques fail to recognize objects in more complex settings e.g., when the object boundary is unknown and cannot be easily found in the image or when the object shape, color and texture exhibit severe changes and cannot be easily described by global measurements extracted from the image. A recent trend consists of selecting local patches (small regions) in the image and describing each patch by a set of local features. This approach does not require a global model for the object and can easily cope with changes in shape, color and texture, achieving good results in very difficult problems [30].

In this paper, we will compare two different strategies for melanoma detection and build a recognition system based on each of them. The first system, describes the dermoscopy image by a set of global features and uses a classifier to discriminate melanomas from non-melanoma lesions (see the block diagram of Fig. 2a). This is a supervised system since the classifier learns to detect the melanoma lesions using a training set of images, labeled by an expert. Each training image, \( I_k \), \( k \in \{1, \ldots, L\} \), is characterized by a feature vector \( x_k \in \mathbb{R}^n \) and by a binary label \( y_k \in \{0, 1\} \). The classifier is trained to discriminate both types of images using the training data \((x_k, y_k), k = 1, \ldots, L\).

The second system, characterizes the dermoscopy image by using a Bag-of-features (BoF) approach (see Fig. 2b) [31], [32], [29], [33], [34]. First, a set of keypoints is selected inside the lesion region. Then each keypoint is characterized by a vector of local features. This feature vector represents color and texture properties in a local patch centered at the keypoint. Since the number of keypoints and local features varies from image to image, we cannot directly feed a classifier with this data. Instead, all local features associated with all the training images are gathered and used to compute a smaller set of prototypes (centroids) denoted as visual words. Then, the local features of each dermoscopy image are assigned to the nearest visual word and an histogram is computed. The histogram counts the number of times each visual word was selected. A statistical classifier is then trained to discriminate melanoma lesions from non-melanoma ones, using the histogram of visual words as input.

The paper tries to determine which of these strategies performs best in melanoma detection. The answer is not simple. If a skin lesion is homogeneous, the first approach will probably be the best since it is able to describe the lesion by a set of global features that use all the available information. However, since skin lesions often have differential structures (pigment network, vascular network, dots, streaks) which are localized and appear in specific regions, these structures may not be well described by global features and local methods may represent them better.

Both systems assume that the dermoscopy image is segmented and the lesion region is separated from healthy skin. Several segmentation methods were proposed for this problem [35], [36], [37], [38], [39]. A comparison among several methods can be found in [40]. However, there is no consensus about which one performs best. We adopted a simple adaptive thresholding technique (see the description of the method in the Appendix) and the output of the segmentation step was verified by an expert and edited when necessary. This procedure was adopted in order to make the system performance independent of the specific algorithm used for segmentation. Otherwise, it would be difficult to interpret the system decisions since errors might be caused by insufficient information provided by the features or by segmentation errors.

IV. GLOBAL LESION CLASSIFICATION SYSTEM

The global classification system consists of three processing steps (see Fig. 2a). First the skin lesion is segmented. Then a vector of global color and texture features is extracted from the skin region. Finally, a statistical classifier is trained using a set of images, labeled by an expert (training set). Three classifiers are considered in this work: AdaBoost classifier [41], Support Vector Machines (SVM) classifier [42] and k-Nearest Neighbor (kNN) classifier.

Concerning the image features, medical doctors use several visual cues such as differential structures, number of colors, symmetry and border transitions. Two strategies will be considered in this paper. In the first case, image features are computed inside the skin lesion, \( R \) (see Fig. 3(b)) while in the second case we split \( R \) into an inner part, \( R_1 \), and the border \( R_2 \) (see Fig. 3(c)). \( R_1 \) is obtained by eroding \( R \) with a disk of radius \( r \), that was empirically chosen to be \( \frac{1}{10} \) of the lesion.
smallest axis after comparing several values. This strategy tries to capitalize on the available knowledge about medical analysis of dermoscopy images: special attention is assigned to the border analysis.

In this paper, we will focus on two types of image features (color and texture) and we will try to assess what role is played by each of them in the final decision: what performance is achieved by color features only, by texture features only and by combining both of them. The features used in this work are presented in the next sections.

![Fig. 3. Original image (a)). Computation of Global features for R b) and R1 + R2 c). Computation of Local features d).](image)

### A. Texture features

Image texture represents the spatial organization of intensity and color in an image, and it can be characterized in many different ways. Some methods use pixel statistics. A classic approach consists of computing the statistics of pairs of neighboring pixels, using the co-occurrence matrix [43]. This idea has been modified and improved in many different ways. For example, local binary patterns (LBP) perform a binary classification of the pixels in the vicinity of each pixel and compute the statistics of the neighboring pixel configurations [44]. Texture has also been described by applying a Fourier transform to the image and by characterizing the spectral energy in different frequency bands [45]. Other image transforms have also been used, e.g., wavelets [46], Laplacian pyramids [47] or linear filters (e.g., Law texture features) [48].

Many works use the gradient histograms to characterize texture and these features have achieved excellent results in several challenging problems. We will use two gradient histograms in this work. First we convert the RGB image into a gray level image by selecting the color channel with the highest entropy [40]. In order to compute the image gradient, we filter the gray level image using a Gaussian filter with $\sigma = 2$, and then compute the gradient vector at each point $g(x) = [g_1(x) \, g_2(x)]^T$ using Sobel masks. The gradient magnitude and orientation are then computed as usual

$$ ||g(x)|| = \sqrt{g_1^2(x) + g_2^2(x)} , \quad \phi(x) = \tan^{-1}\left(\frac{g_2(x)}{g_1(x)}\right). \quad (1) $$

The histograms of the gradient amplitude and orientation are defined as follows,

$$ h_a(i) = \frac{1}{N} \sum_{x \in R} b_i(||g(x)||), \quad i = 1, \ldots, M_a, \quad (2) $$

$$ h_\phi(i) = \frac{1}{N} \sum_{x \in R} \tilde{b}_i(\phi(x)), \quad i = 1, \ldots, M_\phi, \quad (3) $$

where $R$ denotes the set of pixels which were classified as lesion, $N = \#R$ is the number of lesion pixels and $M_a, M_\phi$ are the number of bins used in the amplitude and orientation histograms; $b_i(\cdot), \tilde{b}_i(\cdot)$ are the characteristic functions of the $i$–$\text{th}$ histogram bin, defined by

$$ b_i(a) = \begin{cases} 1 & \text{if } a \text{ belongs to the } i\text{th amplitude bin} \\ 0 & \text{otherwise} \end{cases}, \quad (4) $$

$$ \tilde{b}_i(\phi) = \begin{cases} 1 & \text{if } \phi \text{ belongs to the } i\text{th orientation bin} \\ 0 & \text{otherwise} \end{cases}. \quad (5) $$

When the lesion region $R$ is split into inner region, $R_1$, and border, $R_2$, we compute separate histograms for each of these regions using the same expressions and replacing $R$ by $R_i$ and $N$ by $N_i$.

Fig. 4 shows the histograms of amplitude and orientation for a set of dermoscopy images melanomas (left) and non-melanomas images (right). There is a significant overlap among the features of both classes, in all the histogram bins. This means that the binary decision is not simple.
**B. Color features**

Color information is used by medical doctors in the classification of skin lesions (see Section 1). Therefore, color features have been explicitly used in most CAD systems. The most popular features used in dermoscopy analysis are color statistics such as the mean color and color variance [16], [24]. Most works compute these statistics for the RGB color components. However, other color representations have also been adopted [22].

The RGB color space represents a target color as a mixture of three primary colors: red (R), green (G) and blue (B). The mixture coefficients are denoted as color components. However, this color space has a number of drawbacks: it is not perceptually uniform, it depends on the acquisition setup and exhibits a high correlation among the three color channels [49]. To overcome these difficulties other color representations have been proposed e.g., biologically inspired color spaces such as the opponent color space (Opp) [50], [51], color spaces related to the human description of color such as the hue-saturation and brightness (HSV and HSI) [49], [52] or perceptually uniform color spaces such as the CIE L*a*b* and L*uv [49]. Although there is a 1-1 map between all pairs of color spaces, the choice of the color space modifies the recognition results since their components present different statistics (e.g., some are more correlated than others) and different histograms.

In this paper, we have used all 6 color spaces mentioned above. We characterize the color distribution in the lesion region \( R \) (or \( R_1, R_2 \)) using a set of three color histograms, each of them with \( M_c \) bins. The color histogram associated to the color channel, \( I_c(x) \), \( c \in \{1, 2, 3\} \) is given by

\[
h_c(i) = \frac{1}{N} \sum_{x \in R} b_c(I_c(x)) \quad i = 1, \ldots, M_c, \quad (6)
\]

where \( N = \#R \) denotes the number of pixels in region \( R \) and \( i \) is the histogram bin. The bins are defined by splitting the color component range (which depends on the color space) into \( M_c \) subintervals with the same length.

Fig. 5 shows the concatenation of the histograms \( h_c \) for four different color spaces and splitting the region \( R \) into \( R_1 \) and \( R_2 \). The differences between melanoma and non-melanoma features are more noticeable in these cases.

**V. Detection with Local Features**

BoF methods have recently been used in complex image analysis problems, e.g., image retrieval and object-class recognition in large databases [53], [54]. Since complex objects cannot be represented by a global model, BoF methods represent them by a collection of local models, e.g., image patches. In a first stage, a set of keypoints, \( p_i \in \mathbb{R}^2 \), is defined in the image. In the case of dermoscopy images, we assume that keypoints are nodes of a regular grid of size \( \Delta \times \Delta \) defined in the image domain. Second, each keypoint is described by a feature vector \( x_i \in \mathbb{R}^n \) which represents local information conveyed in the image patch of size \( \Delta \times \Delta \) centered at the \( i-th \) keypoint (see Fig. 3(d)). The size of the patches is equal to the size of the grid to avoid overlapping. Patches for which the area is more than 50% outside the lesion are discarded. Given an input image \( I \) we will denote the family of local features associated to \( I \) by

\[
F = \{x_1, \ldots, x_N\}, \quad x_i \in \mathbb{R}^n, \quad (7)
\]

where \( N \) is the number of keypoints inside the lesion (See Fig. 3, bottom-right).

**A. Training**

Training phase aims to learn the parameters of a BoF classifier using a set of labeled dermoscopy images. Let \( \mathcal{F} = \{f^{(1)}, \ldots, f^{(L)}\} \) be a set of \( L \) images used to train the
classifier and let
\[ \mathcal{F} = \bigcup_{k=1}^{L} F^{(k)}, \]
be the set of all local features detected in the training set. First we will approximate the feature vectors in \( \mathcal{F} \) by a set of prototypes (centroids) \( c_1, \ldots, c_k \). This is done using a clustering algorithm (e.g., k-means algorithm). The prototypes are known as visual words to emphasize the similarity with the bag-of-words method used in text analysis. After obtaining a set of prototypes \( c_1, \ldots, c_k \), all feature vectors in the training set are classified in the nearest prototype and a label is assigned to each feature vector \( x \).

Thus, we have \( I \) as the set of all local features detected in the training set. Since the images in the training set have been classified by an expert, we can train a classifier to predict the labels (melanoma vs. non-melanoma) given the histogram of visual words.

B. Test

In the test phase we wish to classify a new image using the previously learned model. Given an input image \( I \), we compute the keypoints inside the lesion and the corresponding local features. Then we classify the local features using the dictionary of visual words and build the histogram of visual words. Finally, the histogram is classified using the classifier learned during the training phase.

C. Local Features

Each image patch is characterized by color and texture features similar to the ones described in Section IV, i.e., we use gradient histograms and color histograms to characterize the image in each patch.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The systems described in the previous sections were evaluated using a dataset of 176 dermoscopy images (25 melanomas and 151 nevi). Each image was classified by an experienced dermatologist. These images were taken from the database of Hospital Pedro Hispano, Matosinhos, and were obtained by dermatologists during clinical exams using a dermatoscope with a magnification of 20x. They are stored in BMP and JPEG formats and their average resolution is 573x765. The performance of each system was evaluated for each type of feature: T1) gradient magnitude and T2) gradient orientation; C1) RGB, C2) HSV, C3) La*ub*, C4) HSI, C5) L*a*uv and C6) Opp. We have also performed tests using all the color features (C), both texture and color (T+C) and the best texture (Tb) and color (Cb) feature. Since there is always a trade-off between sensibility (SE) and specificity (SP), we evaluate each system configuration using a cost function defined by
\[ \mathcal{C} = \frac{c_{10}(1 - SE) + c_{01}(1 - SP)}{c_{10} + c_{01}}, \]
where \( c_{10} \) is the cost of an incorrectly classified melanoma (False Negative - FN) and \( c_{01} \) is the cost of an incorrectly classified non-melanoma (False Positive - FP). Since we consider that an incorrect classification of a melanoma is a gravest error, we assume that \( c_{10} = 1.5c_{01} \) and \( c_{01} = 1 \) in this paper.

In the following subsections we will address the specific procedures used to evaluate each one of the systems as well as the results obtained.

A. Assessment of the Global Lesion Classification System

The dataset used has a limited number of images. Therefore, we have tested the global system using a leave-one-out approach. This approach requires that for each set of features 176 classifiers are trained, each one using all the images except one, which is used for testing. The final performance measures are computed by averaging the individual measures obtained for each classifier. Class unbalance is another issue that must be addressed. To tackle this problem we have repeated the melanoma features belonging to each training set, until we obtained the same number of examples for both classes. To prevent having equal examples on the training set, we have added Gaussian Noise (\( w \sim N(0, \sigma_0) \)), with \( \sigma = 0.0001 \), to each repeated feature vector.

The performance of the classifier for each feature was evaluated using different combinations of parameters, both in the feature extraction process, by varying the number of bins of the histograms \( n_{b\text{ins}} \in \{15, 16, \ldots, 50\} \), and in the classification process, by performing tests with different classification algorithms (kNN, SVM and AdaBoost). For each classifier we varied specific parameters: namely, the number of neighbors \( k \) and distance \( \text{dist} \) used by the kNN algorithm to compare feature vectors, the type of kernel used by the SVM and the number of combined weak classifiers \( W \) by the AdaBoost. We have considered all the combinations of these parameters and all the possible features and tested more than 30000 different configurations. For each configuration we trained 176 classifiers using the leave-one-out procedure.

The best results obtained using the kNN classifier, selected using the cost criterion (11) are summarized in Table I. We tested all the \( n_{b\text{ins}} \) for different combinations of \( k \in \{5, 7, \ldots, 50\} \). For each \( k \) we used computed the results for three different distances: Euclidean, Kolmogorov and Kullback-Leibler. The parameters which led to the results of Table I can be seen on Table II.

Table I shows that very good results can be achieved using the kNN classifier and each color features outperforms the texture features, which is an interesting result. Another interesting result is that in most of the cases the performance
of the algorithm tends to improve when the lesion region \( R \) is split into two disjoint regions \( (R_1 \text{ and } R_2) \), which suggests that the periphery and inner part of a lesion contain complementary information which allows a better discrimination between melanomas and non-melanomas. This has also been observed in [22]. It is not clear which of the three distances used is best for this specific problem, since all of them are selected and considered to be the best for some of the features. We tried to improve the results by combining the two types of features. However, the system performance decreases when compared to the system performance for the best single color features: \( C_3 \) (SE = 96\% , SP = 79\%) and \( C_6 \) (SE = 100\% and SP = 72\%). This change is explained by the fact that we are increasing the number of features by more than 10 times. The number of training patterns is therefore insufficient to obtain good generalization.

To overcome these difficulties, we have used two additional classifiers which exhibit better generalization performance when the number of features is increased. First, we have tested SVM using two different kernels: linear and a Gaussian Radial Basis Function (RBF). RBF-SVM greatly outperformed the linear one, therefore we will only show results for the second (see Table III).

The SVM classifier does not perform as well as the kNN one. Nonetheless, the obtained results are still interesting and provide some information. As before, the color features outperform the texture ones, which reinforce the idea that this kind of features is more discriminative. In this case, it is not possible to determine if the use of a single region (\( R \)) is better than two \((R_1, R_2)\) since the results are not as clear as before. Despite the low performance of the texture features, we are able to improve the performance of the algorithm by combining \( T_1 \) and \( C_6 \), i.e., the best texture and color features. This feature concatenation leads to the best classification results with SVM: SE = 92\% and SP = 72\%.

AdaBoost has the ability to select appropriate features for a given problem and to train a strong classifier from a combination of weak classifiers [55]. Therefore, this kind of classifier was deemed appropriate for our classification problem since it is able to select a set of informative features for this problem and it is able to discard features which do not improve the classifier performance. AdaBoost was trained using different numbers of weak classifiers \( W \in \{2, 100\} \), achieving the best overall results (SE = 96\% , SP = 80\% for \( C_2/C_4 \) and SE = 92\% , SP = 85\% for \( C_6 \)) with combinations of only 2-5 weak classifiers (see Tables IV and V). Moreover, we also obtained the best global classification results for a single texture feature \( T_1 \) (SE = 96\%, SP = 59\%). Since each weak classifier uses only one bin to classify the data, the experimental results suggest that only a very small number of features is enough to achieve the best scores. This is a remarkable result which suggests that some of the features are much more important than others. It also suggests that best generalization is achieved with a small number of features than with a large feature set.

We sought to determine the impact that increasing the number of weak classifiers (number of selected bins) would have on the classification results. To do so, we varied the value of \( W \), while maintaining the \( n_{\text{max}} \) constant and equal to the one of the best configuration for each single feature (see Table V). The curves obtained are shown on Fig. 6. It is clear that the performance of the classifier depends on the number of features used. This reinforces the idea that most of the bins contain noisy information and have a marginal influence on the classifier output.

The characteristics of AdaBoost allows to explore further the dataset. We performed tests with the best single color features \((C_2, C_4 \text{ and } C_6)\) and their combinations with the best texture feature \((T_1)\) to determine if some histogram bins are more selected than others. This would mean that these bins are very discriminative. The results are shown on Fig. 7. All the
TABLE IV
BEST EXPERIMENTAL RESULTS USING GLOBAL FEATURES AND AN AdaBoost CLASSIFIER.

<table>
<thead>
<tr>
<th>R</th>
<th>R1 + R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>SP</td>
</tr>
<tr>
<td>T1</td>
<td>96%</td>
</tr>
<tr>
<td>T2</td>
<td>80%</td>
</tr>
<tr>
<td>C1</td>
<td>84%</td>
</tr>
<tr>
<td>C2</td>
<td>72%</td>
</tr>
<tr>
<td>C3</td>
<td>88%</td>
</tr>
<tr>
<td>C4</td>
<td>84%</td>
</tr>
<tr>
<td>C5</td>
<td>80%</td>
</tr>
<tr>
<td>C0</td>
<td>96%</td>
</tr>
<tr>
<td>C + T</td>
<td>90%</td>
</tr>
<tr>
<td>T1 + C2</td>
<td>72%</td>
</tr>
<tr>
<td>T1 + C4</td>
<td>84%</td>
</tr>
<tr>
<td>T1 + C6</td>
<td>96%</td>
</tr>
</tbody>
</table>

TABLE V
BEST PARAMETERS FOR THE AdaBoost CLASSIFIER.

<table>
<thead>
<tr>
<th>R</th>
<th>R1 + R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_bins</td>
<td>W</td>
</tr>
<tr>
<td>T1</td>
<td>44</td>
</tr>
<tr>
<td>T2</td>
<td>44</td>
</tr>
<tr>
<td>C1</td>
<td>20</td>
</tr>
<tr>
<td>C2</td>
<td>35</td>
</tr>
<tr>
<td>C3</td>
<td>33</td>
</tr>
<tr>
<td>C4</td>
<td>46</td>
</tr>
<tr>
<td>C5</td>
<td>48</td>
</tr>
<tr>
<td>C6</td>
<td>24</td>
</tr>
<tr>
<td>C</td>
<td>24</td>
</tr>
<tr>
<td>T + C</td>
<td>24</td>
</tr>
<tr>
<td>T1 + C2</td>
<td>33</td>
</tr>
<tr>
<td>T1 + C4</td>
<td>46</td>
</tr>
<tr>
<td>T1 + C6</td>
<td>24</td>
</tr>
</tbody>
</table>

Tests were performed using the best configuration (see Table V) and considering the two regions \( R_1 \) and \( R_2 \), since it is clear that extracting features separately from inner and border improves the classification results. By inspecting the results it is possible to notice that some bins are indeed selected most of the times. These bins correspond to specific channels of each color space, e.g., for \( C_6 \) the two most selected bins correspond to bins of the histogram of the second channel (\( O_2 \)) and for \( C_3 \) and \( C_4 \) the most selected bin is from the histogram of the second channel (\( S \) - Saturation), while the two other correspond to the third channel (\( V \) - Value and \( I \) - Intensity, respectively). An interesting result is that when a fusion of two best features is performed the classifier only selects bins from color histograms, which demonstrates that color features have more discriminative information. The classifier selects information from \( R_1 \) and \( R_2 \), which suggests that both regions are informative. We have computed the mean and median values for the two most selected bins for each of the three best color features (see Table VI). The most selected bin is labeled as “1” and the second most selected bin as “2” (recall Fig. 7, to identify the corresponding bins for each color feature). These bins have values considerably different for the two classification classes, which demonstrates that these bins can be used to distinguish between melanomas and non-melanomas.

Table VI
MEAN AND MEDIAN VALUES FOR THE TWO MOST SELECTED BINS (THE MOST SELECTED BIN IS “1” AND THE SECOND MOST SELECTED IS “2”) FOR \( C_2 \), \( C_4 \) AND \( C_6 \).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Mean</th>
<th>Median</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.027</td>
<td>0.025</td>
<td>0.034</td>
<td>0.028</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>Non-Melanoma</td>
<td>0.005</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0.010</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.002</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.002</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0.010</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 8 shows 4 examples classified by the proposed system using Adaboost, \( C_6 \) and the best combination of parameters for this feature. These examples correspond to the four possible cases: True Negative (TN) - correctly classified non-melanoma; False Positive (FP) - incorrectly classified non-melanoma; False Negative (FN) - incorrectly classified melanoma and True Positive (TP) - correctly classified melanoma.

Although we do not use the same features and datasets as the other works found on literature, a tentative comparison can be performed. Faziloglu et al. [26] report a \( SE = 84.3\% \), \( SP = 83.0\% \) and Stanley et al. [23] report the following values: \( SE = 87.7\% \), \( SP = 74.9\% \). Both these groups use only color features and their datasets are larger than ours (258 and 226 images, respectively), possessing an equal proportion of melanomas and non-melanomas. With our best configurations and classifier we achieve results similar to the ones reported by these two groups: \( SE = 96.0\% \), \( SP = 80.0\% \) using \( C_2/C_4 \) and \( SE = 92.0\% \), \( SP = 85\% \) using \( C_6 \). Our classification results are also similar to the ones reported by Seidenari et al. [21] (\( SE = 87.5\% \), \( SP = 85.7\% \)). The performance achieved using the best texture feature \( T_1 \) (64.2\% for \( SE = 96.0\% \), \( SP = 59\% \)) is similar to the ones reported by Yuan et al. [27] (accuracy of 70.0\%).
B. Assessment of the Local Lesion Classification System

A similar procedure was used to optimize and test the CAD system based on local features. The same data base of images was used in this case. However, the algorithm was evaluated using stratified 10-fold cross validation, since the increase of computational time associated with the local methods turned leave-one-out evaluation unfeasible. The melanoma images was evenly distributed by all the folders and their features was replicated as before, in order to balance both classes.

Since the feature extraction operation and the BoF classifier (see Fig. 2b) depend on a number of parameters, we have explicitly varied each of them in order to achieve the best classification scores. In the case of features we changed the grid size $\Delta \in \{20, 40, \ldots, 100\}$ and the number of histogram bins in the range $n_{\text{bins}} \in \{15, 25, 35, 45\}$. The visual words are obtained by applying the k-means clustering algorithm to the training vector features. The number of visual words was varied in the range $K \in \{100, 200, 300\}$. The test image is classified by the kNN algorithm as being melanoma or non-melanoma. Three distances were considered (Euclidean, Kolmogorov, Kullback-Leibler) and the number of neighbors was varied in the range $k \in \{5, 7, \ldots, 25\}$. We have tested all the possible combinations of these 5 parameters for each type of feature. This amounted to 44928 configurations, each of them involving the training of 10 classifiers, using the stratified 10-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>SE</th>
<th>SP</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>33%</td>
<td>76%</td>
<td>0.495</td>
</tr>
<tr>
<td>$T_2$</td>
<td>90%</td>
<td>45%</td>
<td>0.281</td>
</tr>
<tr>
<td>$C_1$</td>
<td>100%</td>
<td>72%</td>
<td>0.110</td>
</tr>
<tr>
<td>$C_2$</td>
<td>87%</td>
<td>88%</td>
<td>0.128</td>
</tr>
<tr>
<td>$C_3$</td>
<td>93%</td>
<td>85%</td>
<td>0.104</td>
</tr>
<tr>
<td>$C_4$</td>
<td>95%</td>
<td>79%</td>
<td>0.115</td>
</tr>
<tr>
<td>$C_5$</td>
<td>100%</td>
<td>75%</td>
<td>0.099</td>
</tr>
<tr>
<td>$C_6$</td>
<td>93%</td>
<td>68%</td>
<td>0.172</td>
</tr>
<tr>
<td>$T + C_1$</td>
<td>100%</td>
<td>70%</td>
<td>0.118</td>
</tr>
<tr>
<td>$T_1 + C_3$</td>
<td>91%</td>
<td>86%</td>
<td>0.114</td>
</tr>
<tr>
<td>$T_1 + C_5$</td>
<td>98%</td>
<td>79%</td>
<td>0.100</td>
</tr>
</tbody>
</table>

Tables VII and VIII summarize the results achieved in this
study. Tables VII show the best results achieved by the BoF classifier for each kind of features and Table VIII shows the parameters of the best configurations. In the case of texture features, the best results are obtained with the histogram of orientation $T_2$. The role of the orientation seems to be much more important when computed in local patches than when it was obtained from the whole lesion. Color features achieve better results as before, leading to remarkable scores, e.g., $SE = 100\%$, $SP = 75\%$ or $SE = 93\%$, $SP = 85\%$. The concatenation of texture and color features lead to excellent results ($SE = 98\%$, $SP = 79\%$) but it does not improve the best color feature configuration.

Table VIII also allows to draw interesting conclusions. The best distance for these features is the Kullback-Leibler divergence. This is not surprising since all the features considered in this study are normalized histograms. The best block size is 80 for most features considered in these experiments. This is a bit surprising since we would expect that smaller block sizes would allow a better representation of local differential structures (e.g., pigment network, dots, streaks, blue whitish veil). These differential structures are difficult to characterize even when the analysis is made by medical experts trained in dermoscopy and may be insufficiently described by the automatically extracted features.

![Image](image_url)

**TABLE VIII**

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$n_{bins}$</th>
<th>$K$</th>
<th>dist</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>60</td>
<td>15</td>
<td>200</td>
<td>9</td>
</tr>
<tr>
<td>$T_2$</td>
<td>80</td>
<td>25</td>
<td>300</td>
<td>25</td>
</tr>
<tr>
<td>$C_1$</td>
<td>80</td>
<td>25</td>
<td>300</td>
<td>13</td>
</tr>
<tr>
<td>$C_2$</td>
<td>80</td>
<td>25</td>
<td>300</td>
<td>25</td>
</tr>
<tr>
<td>$C_3$</td>
<td>80</td>
<td>25</td>
<td>300</td>
<td>19</td>
</tr>
<tr>
<td>$C_4$</td>
<td>80</td>
<td>25</td>
<td>300</td>
<td>13</td>
</tr>
<tr>
<td>$C_5$</td>
<td>80</td>
<td>15</td>
<td>300</td>
<td>13</td>
</tr>
<tr>
<td>$T + C$</td>
<td>80</td>
<td>25</td>
<td>200</td>
<td>17</td>
</tr>
<tr>
<td>$C$</td>
<td>80</td>
<td>25</td>
<td>200</td>
<td>15</td>
</tr>
<tr>
<td>$T_2 + C_3$</td>
<td>60</td>
<td>25</td>
<td>300</td>
<td>15</td>
</tr>
<tr>
<td>$T_2 + C_5$</td>
<td>100</td>
<td>15</td>
<td>300</td>
<td>13</td>
</tr>
</tbody>
</table>

Fig. 9 shows the variations of SE and SP for the best single color and texture features ($C_3$, $C_5$ and $T_2$), using different values of $K$ (1st row), $\Delta$ (2nd row), $n_{bins}$ (3rd row) and the three possible distances (4th row). The values that were kept unchangeable are the ones of the best configuration for each single feature (see Table VIII). It is clear that the number of visual words $K$ influences the performance of the classifier. Furthermore, the size of the grid $\Delta$ seems to also have a great impact on the performance of the classifier.

Although most of previously published systems use global features, we can compare these results with the ones presented in [25] which achieved $SE = 80\%$, $SP = 78\%$ in database of 100 dermoscopy images with 30 melanomas. We stress that these results were obtained using texture features only (Gabor filters and wavelets) with a small block size ($\Delta = 16$).

**C. Comparison between global and local classification systems**

Comparing local and global strategies for the detection of melanomas (Tables IV and VII) we conclude that both strategies achieve very good results in this database. The local features with BoF achieves slightly better results. In both cases, the color features perform significantly better than texture features. This is a bit surprising since differential structures play an important role in medical diagnosis of dermoscopy images and therefore, we would expect that this behavior could be captured by texture descriptors.

The computational time for feature extraction and classification of an image using the two best systems can be seen on Table IX. Computing local features takes more time than computing global ones, this was expected since in the case of
the local system features are extracted for each patch, whilst in the case of a global system the features are only extracted twice (the shown values are for the case of splitting the lesion in inner and border).

<table>
<thead>
<tr>
<th>Feature Extraction</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local System</td>
<td>Global System</td>
</tr>
<tr>
<td>2 Regions - 0.13s</td>
<td>AdaBoost - 0.48s</td>
</tr>
<tr>
<td>0.39s</td>
<td>Histogram+kNN - 0.022s</td>
</tr>
</tbody>
</table>

VII. CONCLUSIONS

This paper compares two different strategies for the detection of melanomas in dermoscopy images based on local and global features. Most previous works use global features (texture, shape, color) associated to the whole lesion followed by a binary classifier trained from the data. However, local features are gaining increasing importance in many image analysis problems (e.g., image retrieval, object recognition) and may be seen as a promising alternative.

A second driving idea associated with this paper is the evaluation of the role played by color and texture features in the decision. Most works consider both types of features together but they do not attempt to clarify if one of these features play a more relevant role. This question is explicitly addressed in this paper.

Both issues were tested using a data set of dermoscopy images from Hospital Pedro Hispano and the system parameters were tuned by exhaustive testing of many thousands of classifiers. It was concluded that color features perform much better than texture features alone achieving classification scores of $SE = 93\%$, $SP = 85\%$. Concerning the use of global vs local features and classification strategies, both systems achieve very good results. The best global system results were achieved when the lesion region was split into two sub-regions (inner and border) leading to $SE = 96\%$, $SP = 80\%$. However, the local system performs slightly better in terms of the classification cost used in this study ($SE = 100\%$, $SP = 75\%$).

APPENDIX

LESION SEGMENTATION

Lesion segmentation is obtained using an adaptive thresholding algorithm which comprises three steps: i) histogram computation, ii) peak detection and iii) threshold estimation. These three steps are described in the sequel.

First, the input image is converted into a gray level image by selecting the channel with the highest entropy [40]. Then, the histogram of intensity, $h(i), i = 0, \ldots, 255$, is computed using all the pixels inside a rectangular window, located at the center of the image. The histogram is then lowpass filtered with a Gaussian filter. The filter impulse response is $G(i) = C \exp\left(-i^2/(2\sigma^2)\right)$, where $\sigma = 5$ and $C$ is a normalization constant.

We assume that the analysis intersects the lesion, and one of the histogram modes is associated to the lesion. We therefore extract the most significant peaks of the histogram (with a maximum 2). If the histogram has a single peak located at $i_1$, the threshold is defined by $T = i_1 + \Delta T$. The $\Delta T$ offset was empirically tuned and set to 15. If two significant peaks are extracted, we choose the threshold which corresponds to the maximum valley. The valley is defined as follows: given two peaks located at $i_1, i_2$, $i_2 > i_1$, we define the valley at $\Delta i$ as the difference between the straight line defined by the peaks and the histogram amplitude at $i$ (see Fig. 10).

The only information missing is how to select the most significant peaks in a histogram if more than two peaks are detected. This involves two steps. First close peaks are merged and replaced by a single peak: the largest one. We consider that two peaks located at $i_1, i_2$ are close if $|i_1 - i_2| < \Delta i$ ($\Delta i = 8$). If the number of peaks is larger than 2 after this step, the highest peak is selected. Then, all the remaining ones are tested. The second selected peak is the one which corresponds to the largest depth of the valley between two histogram peaks.

REFERENCES


