A Clinically Oriented System for Melanoma Diagnosis Using a Color Representation

Catarina Barata [1], M. Emre Celebi [2] and Jorge S. Marques [1]

Abstract— **Computer Aided-Diagnosis (CAD) systems have been proposed to help dermatologists diagnose melanomas. However, these systems fail to provide a medical explanation for the diagnosis. This makes the dermatologists unsure about their use, since they are not easy to understand. In this paper we propose a CAD system that extracts a clinically inspired color description of the lesion and then, uses this information to discriminate melanomas from benign lesions. The proposed system is also capable of showing the extracted color features, making the system and its decisions more comprehensible for practitioners. The development of this system is hampered by the lack of a database of detailed annotate dermoscopy images. Nonetheless, we are able to tackle this issue using an image annotation framework based on the Correspondence-LDA algorithm. This method is applied with success to the identification of relevant colors in dermoscopy images, obtaining an average Precision of 84.9% and a Recall of 85.5%. The proposed color representation is then used to classify skin lesions, resulting in a Sensitivity of 78.9% and Specificity of 76.7%, these values are promising and comparable with the state-of-the art.**

I. INTRODUCTION

Dermatologists diagnose melanomas in dermoscopy images following two steps. First, they extract clinical features from the image (e.g., main colors, differential structures such as dots, streaks or pigmented network). Then, they classify the image based on these features [1], [12]. Several Computer Aided-Diagnosis (CAD) systems were developed to assist dermatologists in their routine practice [9]. However, most of these systems extract abstract features that have no medical meaning. The system gives a final decision (e.g., melanoma vs. benign) but no clinical information is shown to the doctors allowing them to understand why the decision was taken and whether it makes sense or not. We need a different type of systems designed to work in a collaborative way with practitioners, providing not only a diagnosis but also medical information that may help dermatologists understand the systems decision [7].

The goal of this paper is to propose a CAD system that first extracts clinically inspired features and then, uses this information to discriminate melanomas from benign lesions. The developed system is also capable of showing dermatologist the extracted features in a comprehensive way. More precisely, the system simultaneously provides a text description of the detected medical criteria and it is capable of highlighting them inside the lesion. We expect that by providing this information, we are able to gain the trust of dermatologists regarding the decisions of the system. The development of such a system poses a great challenge since we cannot rely on large databases of images with detailed medical annotations. Most databases available provide text annotations concerning color, or differential structures but do not provide the location of such structures in the images (e.g., the EDRA database [1]). Therefore, if we want to work in a realistic setup, the CAD system must be trained from weakly annotated images (with text labels only). This makes the problem difficult.

This paper proposes a CAD system for melanoma diagnosis that provides detailed clinical information about the presence and spatial distribution of colors. We adopt the probabilistic model Correspondence-LDA (corr-LDA) that was proposed in the context of image interpretation [4]. This model learns the joint probability distribution between medical labels and image regions from weakly annotated data (text labels) and allows to successfully retrieve text and visual color information from dermoscopic images. Since the location of each color is retrieved, it can be shown to the medical doctors. In a second step, the estimated colors are used to perform a global classification of the lesion in order to discriminate melanomas from benign nevi. The first stage of this system was recently proposed by us in [2] and it is reviewed here for the sake of completeness. We extend this stage with a decision stage that is able to provide a global diagnosis from the local features.

II. CAD SYSTEM

Fig. 1 shows the block diagram of the proposed CAD system based on color information, as well as an example of the desired output. Our goals are the following: i) obtain a medical color representation of the lesion, and ii) use this information to obtain a diagnosis (melanoma or benign). The system must be also capable of providing information to dermatologists, namely associate image regions with one of six possible clinically relevant colors (dark and light brown, red, white, black, and blue-gray [12]) and provide a set of text labels $\mathbf{w}$ to describe the colors.

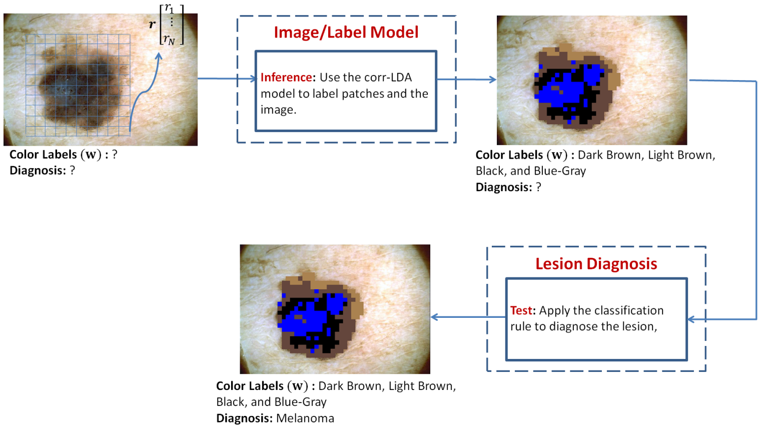
The proposed system works as follows. First, the dermoscopy image is partitioned into N small square patches of dimension 12×12 pixels, using a regular grid. Patches containing less than 50% lesion pixels are discarded. Then, each patch is characterized by a feature vector $r_n, n = 1, ..., N$. Based on previous works, we decided to use the mean color vector in the HSV color space to characterize the

*This work was funded by grant SFRH/BD/84658/2012 and by the FCT project FCT [UID/EEA/50009/2013]

[1] C. Barata and J.S. Marques are with Institute for Systems and Robotics, Instituto Superior Tecnico, Lisboa, Portugal ana.c.fidalgo.barata@ist.utl.pt

[2] Department of Computer Science, Louisiana State University, Shreveport LA, USA

978-1-4244-9270-1/15/$31.00 ©2015 IEEE



r
Image/Label Model
Inference: Use the corr-LDA model to label patches and the image.
Color Labels (w) : ?
Diagnosis: ?
Color Labels (w) : Dark Brown, Light Brown, Black, and Blue-Gray
Diagnosis: ?
Lesion Diagnosis
Test: Apply the classification rule to diagnose the lesion,
Color Labels (w) : Dark Brown, Light Brown, Black, and Blue-Gray
Diagnosis: Melanoma

Fig. 1. Proposed CAD System.

patches [3]. The set of all the feature vectors is denoted $\mathbf{r} = \{r_1, r_2, ..., r_N\}$.

The "Image/Label Model" block performs color detection and image annotation. Image annotation consists of finding a set of text labels $\mathbf{w} = \{w_1, ..., w_M\}$, $M \in \{1, 2, ..., 6\}$ that identifies the colors present in the lesion, as well as associating them with each of the image patches. This is exemplified in Fig. 1. Annotation is achieved by inference using the corr-LDA algorithm [4]. This algorithm is capable of computing the following probabilities of interest: i) $p(w_m|r_n)$, which is distribution of a color label given a single patch, used to perform patch labeling; and ii) $p(w_m|\mathbf{r})$ - distribution of a color label given the entire image, used to obtain the text labels. We discuss these issues in Section III-A.

The second block performs lesion diagnosis. In this step, we use the color information extracted using corr-LDA to obtain a decision about the type of lesion. In Section IV we address the classification strategies considered in this work.

III. CORR-LDA AND COLOR DETECTION

A. corr-LDA

corr-LDA is a generative model used for image annotation [4]. This algorithm assumes that images and their respective captions are generated in a sequential way. First, a set of N feature vectors $\mathbf{r} = \{r_1, ..., r_N\}$ is generated characterizing all of the image patches. Each of these descriptors r_n is generated conditioned on a hidden variable (topic) z_n; $\mathbf{z} = \{z_1, ..., z_N\}$ being the set of topics that was used to obtain the image. Finally M text labels are generated as follows. For each annotation, one of the image patches is selected and a corresponding annotation w_m is drawn conditioned on the topic that was used to generate the patch descriptor [4].

Each of the previous variables is generated using a parametric distribution, summarized as follows:

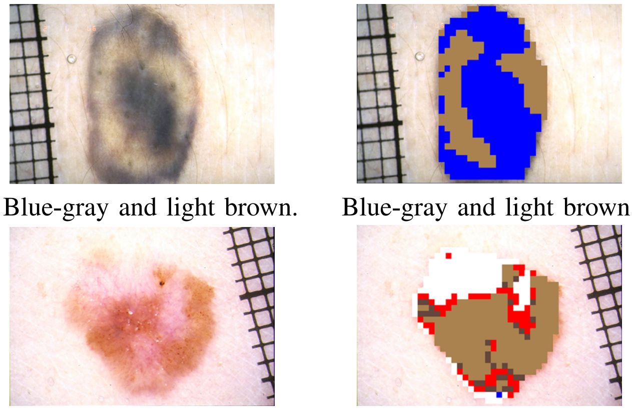
1) For each image d, from a set of D images, sample a topic distribution $\theta \sim \text{Dirichlet}(\alpha)$.
2) For each of the N image patches r_n
   a) Sample $z_n \sim \text{Multinomial}(\theta)$.
   b) Sample $r_n \sim p(r|z_n, \Omega)$ from a von-Mises/Gaussian distribution conditioned on z_n.
3) For each of the M labels w_m
   a) Sample $y_m \sim \text{Uniform}(1, ..., N)$.
   b) Sample $w_m \sim p(w|y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the z_{y_m} topic.

Here, α is the Dirichlet parameter and has the dimension of the number of topics (K). Ω is the set of parameters of one of the $k = 1, ..., K$ von-Mises/Gaussian distributions [2] that characterize the image patches, and β is the distribution of the possible labels over each of the k topics. These are model parameters, while θ is an image specific parameter that equals K and is sampled once per image. y_m is a latent indexing variable that takes values between 1 and N^d and is used to select the patch that generates the m-th annotation.

1) Inference: The inference problem associated with corr-LDA requires the computation of the posterior distribution of the latent variables $(\theta, \mathbf{z}, \mathbf{y})$ given the observations (patch features and annotations). Unfortunately, an exact computation of this posterior is not possible. An approximation can be estimated using Variational Inference [5], [4]. This strategy consists of applying the Jensen's Inequality to find a family of lower bounds of the log-likelihood. The lower bounds are indexed to a set of variational parameters (γ, ϕ, λ) that are unique for each image. The optimal variational parameters are the ones that minimize the Kullback-Leibler divergence between the approximation and the true posterior.

2) Parameter Estimation: Given a set of training pairs of features/annotations $(\mathbf{r}^d, \mathbf{w}^d)$, $d = 1, .., D$, our goal is to obtain the maximum likelihood estimates of the model parameters (α, β, Ω), which characterize the training database. These estimates can be obtained using a variational Expectation-Maximization (EM) method that maximizes the aforementioned lower bound. More specifically, this process consists of iteratively applying the following two steps until convergence

- **E-Step:** The variational parameters $(\gamma^d, \phi^d, \lambda^d)$ are estimated for each image d in the dataset and the lower bound is computed.
- **M-Step:** The model parameters α, β, and Ω are estimated by maximizing the lower bound obtained in the E-step.



Blue-gray and light brown. Blue-gray and light brown.

Dark and light browns, red, and white. Dark and Light browns, red, black, and white.

Fig. 2. Original image medical labels (left) and output of corr-LDA (right).

Due to space constraints, please refer to [2] to find the update equations of the variational and model parameters.

B. *Color Detection*

In order to perform color detection using corr-LDA we start by dividing the dermoscopy images into small patches and characterize them as described in Section II. The remaining steps can be divided into two phases: training and testing.

1) Training: We use a set of D dermoscopy images that have been labeled by an expert to estimate the model parameters (α,β,Ω), as described in III-A.2. Since the annotations provided by the dermatologists are strings, it is necessary to convert it into a binary vector $\mathbf{w}$ of length $M = 6$ (same as the number of colors [12]) where $w_m = 1$ if the m-th color is present and 0 otherwise.

2) Testing: The annotation of new images is performed as follows. First we apply the E-step to each of the images in order to determine their corresponding variational parameters. Then we compute the probabilities $p(w_m|r_n)$ and $p(w_m|\mathbf{r})$ as described in [4], [2], in order to obtain respectively the patch and image annotations. The text labels are obtained by comparing $p(w_m|\mathbf{r})$ with an empirically determined threshold, validating those colors that are above the threshold. Examples of the patch labeling and image annotation processes can be seen in Fig. 2.

IV. LESION DIAGNOSIS

corr-LDA allows us to obtain local (patch) and global (image) color labels for the lesions. In this step we want to convert these medical color annotations into an appropriate description that can be used to discriminate melanomas from benign lesions. Since we do not know the optimal way to describe the lesions, we investigate four different strategies:

- **Number of Colors (i):** This is the simplest and most clinically oriented description. We simply count the number of global labels (colors) that are obtained for a given lesion and use this number to characterize the lesion.
- **Present/Absent Colors (ii):** Instead of counting the number of colors, we can describe the lesion stating which are the colors that are present or absent. We represent the lesion by a feature vector $\mathbf{c}^d$ of length 6, where c_m^d is equal to 1 if the m-th color is present and 0 otherwise. The reader might identify this description as the same one that we use to represent the medical color annotations during the train of corr-LDA.
- **Distribution of Color Annotations (iii):** Another possibility is to describe the images using the conditional distribution $p(w|\mathbf{r})$, which provides the probability of each color in the lesion. We represent each lesion by a feature vector $\mathbf{c}^d$ of size 6, where $c_m^d = p(w_m|\mathbf{r}^d)$ and m identifies one of the six colors.
- **Number of Patches per Topic (iv):** The variational parameters γ_k^d approximately correspond to the k-th model parameter α_k plus the expected number of patch features that were generated by the k-th topic [5]. In [5] it was proposed that the number of patches per topic could be used as a descriptor. Thus, we also test this hypothesis. The feature vector $\mathbf{c}^d$ obtained in this case has the same length as the number of topics and $c_k^d = \alpha_k - \gamma_k^d$.

Each of the aforementioned descriptors is used to classify the lesions as melanoma or benign. The classification method based on feature (i) is the simplest one. We classify the lesion as melanoma if the number of annotations/colors is higher than 3. This threshold is defined based on the findings of MacKie et al. [10]. The diagnosis based on the remaining descriptors requires the use of a classification algorithm. This means that we have to train a classifier using a training set of images previously diagnosed by an expert. Then, the obtained classification rule is used to classify new lesions as melanoma or benign. The classifier considered in this work is AdaBoost [8].

V. EXPERIMENTAL RESULTS

The experiments were performed using a dataset of 482 dermoscopy images (50% melanomas) randomly selected from the commercial database EDRA [1]. This is a multi-source database that contains dermoscopy images from three different university hospitals: University Federico II of Naples (Italy), University of Graz (Austria), and University of Florence (Italy). Each of the lesions has been analyzed by a group of experienced dermatologists and text color annotations were available for 344 out of the 482 images. This reduced set was used to train and evaluate the color detection method based on corr-LDA while the full set was used to train and test the automatic lesion diagnosis.

To evaluate the performance of corr-LDA in the color detection problem we compute two metrics for each color: Precision and Recall. Precision corresponds to the proportion of images where a specific color was correctly annotated among all the images where that color was detected. Recall is the percentage of images where the color was correctly annotated. The performance of lesion diagnosis is evaluated using the metrics Sensitivity and Specificity. The aforementioned

TABLE I
COLOR DETECTION RESULTS.

	#Images	Precision	Recall
Blue-Gray	226	91.2%	86.7%
Dark-Brown	303	94.1%	94.1%
Light-Brown	247	89.6%	90.7%
Black	179	82.7%	85.5%
Red	32	90.9%	62.5%
White	15	60.9%	93.3%
Average	-	84.9%	85.5%

TABLE II
LESION DIAGNOSIS RESULTS.

	Threshold		**AdaBoost**	
Feature	SE	SP	SE	SP
(i)	50.6%	87.1%	-	-
(ii)	-	-	83.8%	50.1%
(iii)	-	-	64.8%	64.7%
(iv)	-	-	**78.9%**	**76.7%**

metrics were computed using a 10-fold cross validation approach in which the dataset is divided into ten subsets, each with approximately the same number of melanomas and benign lesions. We used the same folds to train and test the color detection and lesion diagnosis blocks. Therefore, ensuring that the reduced set of 344 images was fairly split among the 10 folds, such that we had enough images for train and test each time.

The performance of color detection is shown in Table I. This table shows the scores obtained for each color as well as the average performance of the probabilistic model. Despite the difficulty of color detection based on color labels, these results show that corr-LDA performs well.

Table II shows the classification scores obtained using each of the features described in Section IV. These results show that each feature performs differently and that some of them are more appropriate to identify melanomas than others.

The number of colors (feature (i)) was based on the findings of MacKie et al. [10]. They found that the presence of more than 3 colors was a sign of malignancy, obtaining a SE = 92% and a SP = 51% on their experiments. Applying the same strategy to our database lead to different results, with a significantly lower SE (50.6%) an higher SP (87.1%).

Assessing the colors that can be found in the lesion (feature (ii)) seems to be highly specific of melanoma, but results in a large number of incorrectly diagnosed benign lesions. Feature (iii), that corresponds to the distribution $p(w|\mathbf{r})$, does not allow a good discrimination between the melanoma and benign classes. The best trade-off between SE and SP is obtained using feature (iv), i.e., the number of regions per topic. We achieve the best classification results with: SE=78.9% and a SP = 76.7%.

These results are comparable with those of other works where color information is used to diagnose melanomas. Seidenari et al. report scores of SE=69.9% and SP=85.8% on calibrated image data, while Celebi and Zornberg [6] report SE=61.6% SP = 75.8% on uncalibrated image data. The SE obtained in our work is higher than that of other works (78.9%) while SP is lower than in the case of [11] and similar to [6]. Overall, our method obtains the best trade off between SE and SP and we stress that these results are obtained using weakly annotated images.

VI. CONCLUSIONS

We proposed a CAD system trained from weakly annotated data that is able to identify clinically relevant colors and use this information to diagnose the lesions. The described system also provides a clinical description of the image including text and visual information. The visual information is quite significant since it allows dermatologists to understand the decisions made by the system. This way, the dermatologist is able to evaluate the performance of the system and decide if he trusts it or not. The results were promising, with the following average scores for color detection: Precision = 84.9% and Recall=85.5%. These scores were obtained using the corr-LDA algorithm. We have investigated different strategies for converting the color information into a melanoma descriptor. The best results led to a Sensitivity of 78.9% and a Specificity of 76.7%.

It is important to stress that a medical diagnosis is performed based on more criteria besides color and basing the decision solely on this criterion could lead to an incorrect decision. In future work we would like to extend our corr-LDA model to other dermoscopic structures and use that information to improve the CAD system results.

REFERENCES

[1] G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, V. Hofmann-Wellenhog, D. Massi, G. Mazzocchetti, M. Scalvenzi, and I. H. Wolf, *Interactive Atlas of Dermoscopy*. EDRA Medical Publishing & New Media, 2000.

[2] C. Barata, M. E. Celebi, and J. S. Marques, "Color detection in dermoscopy images based on scarce annotations," in *to be presented at IbPRIA*, 2015.

[3] C. Barata, M. A. T. Figueiredo, M. E. Celebi, and J. S. Marques, "Color identification in dermoscopy images using gaussian mixture models," in *ICASSP'14*, 2014, pp. 3611–3615.

[4] D. Blei and M. Jordan, "Modeling annotated data," in *26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.

[5] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[6] M. E. Celebi and A. Zornberg, "Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification," *IEEE Systems Journal*, vol. 8, no. 3, pp. 980–984, 2014.

[7] S. Dreiseitl and M. Binder, "Do physicians value decision support? a look at the effect of decision support systems on physician opinion," *Artificial intelligence in medicine*, vol. 33, no. 1, pp. 25–30, 2005.

[8] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[9] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: a review," *Artificial intelligence in medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[10] R. MacKie, C. Fleming, A. McMahon, and P. Jarrett, "The use of the dermatoscope to identify early melanoma using the three-colour test," *British Journal of Dermatology*, vol. 146, no. 3, pp. 481–484, 2002.

[11] S. Seidenari, G. Pellacani, and C. Grana, "Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment," *British Journal of Dermatology*, vol. 149, no. 3, pp. 523–529, 2003.

[12] W. Stolz, A. Riemann, and A. B. Cognetta, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *European Journal of Dermatology*, vol. 4, pp. 521–527, 1994.