A System for the Analysis of Dermoscopy Images Using Weak Annotations

Catarina Barata[1]
ana.c.fidalgo.barata@ist.utl.pt
M. Emre Celebi[2]
ecelebi@uca.edu
Jorge Marques[1]
jsm@isr.tecnico.ulisboa.pt

[1] Institute for Systems and Robotics
Instituto Superior Técnico
Lisboa, Portugal

[2] Department of Computer Science
University of Central Arkansas
Arkansas, USA

Abstract

This paper proposes a two-step approach for the analysis of dermoscopy images. In the first step, we detected dermoscopic criteria (structures and colors), which are used by dermatologists in their medical analysis. In the second step, this information is used to automatically diagnose skin cancer.

The extraction of dermoscopic criteria from skin lesions is a challenging task because the amount of detailed annotated images is scarce. We solve this task by using a probabilistic model (topic model) learned from weakly annotated data. This approach overcomes the need for completely annotated datasets, only requiring text labels. The second step uses the detected criteria to train a Random Forest classifier. The system achieves a good classification score: sensitivity of 85.8% and a specificity of 71.1%. Nonetheless, the main advantage of this system with respect to others is its ability to justify the decision based on medical criteria.

1 Introduction

Dermoscopy is a popular modality used by dermatologists to obtain magnified images of skin lesions. Over the last years, there has been an increasing interest in the development of automatic systems for the analysis of these images in order to detect skin cancer, namely melanoma. Most systems follow a standard classification framework: i) lesion segmentation; ii) feature extraction, and iii) classification. Although the performance of these systems is good, dermatologists do not trust these systems because: i) they are black boxes and the extracted features have no medical meaning and ii) the number of wrong decisions is still too high.

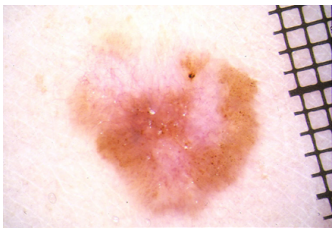
This paper aims at extracting features similar to those adopted by dermatologists in their daily procedures. Two examples are the ABCD rule [9] and the 7 point checklist [1] methods that identify specific properties inside the lesion region. These can be either relevant structures such as pigment network, dots/globules, streaks, and blue whitish veil, or specific colors, such as the six different colors considered in the ABCD rule (dark and light brown, black, white, red, and blue). The presence of high number of colors and structures in a lesion is interpreted as a malignancy cue.

The automatic extraction of such criteria is challenging. There are many attempts to solve this problem but most works only focus on a single criterion [8] and use private databases specific for that criterion. To the best of our knowledge, it is not possible to find a public database that comprises segmented images of the medical criteria.

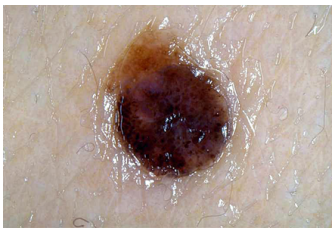
To overcome this limitation, we solve the criteria detection problem using weakly annotated data, as exemplified in Figure 1. Instead of using segmented images for each of the criterion we train a probabilistic model using text information, which states if the criterion is present or not in the image (without specifying its location). This can be done using a generative model of the image based on latent variables. The model used in this work is called Correspondence Latent Dirichelet Allocation (corr-LDA) [4, 5]. Although this model has been used with success in image annotation tasks performed in large databases, it was only recently used in the context of dermoscopy [3].

Figure 2 shows a typical input image and the output of the proposed system displaying the detected colors and structures. After extracting the medical features from the image, the system trains a support vector machine to classify new images as melanoma or benign.

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



Colors: Dark brown, light brown, red, and white.
Structures: Dot and regression. areas.
Diagnosis: Melanoma.



Colors: Dark brown, black, light brown, and blue.
Structures: Pigment network, dots, and blue-whitish veil.
Diagnosis: Benign.

Figure 1: Images and annotations performed by dermatologists [2].

2 Proposed System

We wish to automatically identify the presence and location of several medical criteria organized into two groups: colors (dark and light brown, black, white, red, and blue) and structures (pigment network, dots, blue whitish veil, and regression areas.)

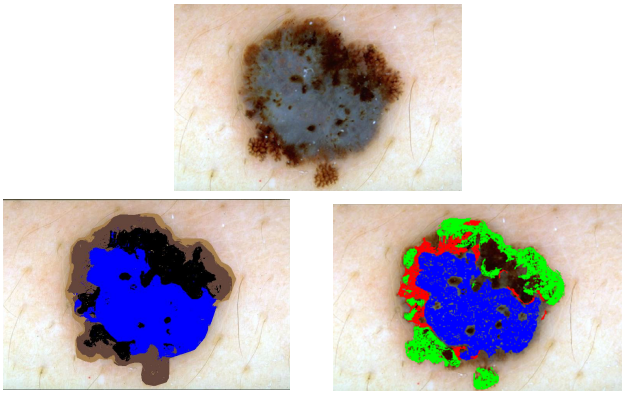
The training set consists of 804 images from EDRA database [2], each of them weakly annotated by a group of experts, *i.e.*, for each image there is a set of binary labels stating whether each criterion is present (see examples in Figure 1). On a first stage, each image is split into a tentative set if homogeneous regions, regarding color and texture. This is accomplished using the superpixel algorithm proposed in [6]. The n-th region is then characterized by a set of image features r_n (color and texture).

Since we want to locate each medical criteria in the image, we assume that there is a local label associated to each region. However, the information provided by the experts is a global label $w \in \{0,1\}$ valid for the entire image. Therefore, we need to find the relationship between local (region) labels and the global (image) one.

A strategy to address the previous problem consists of defining a joint probability distribution, $p(\mathbf{r},w)$, of the observed region features $\mathbf{r} = \{r_1,\ldots,r_N\}$ and image label w. One possible method to estimate this probability is the corr-LDA generative model, which uses latent variables, called topics, to increase its flexibility [4]. For each region n, the model performs three generative sequential processes, *i.e.*, it generates three variables. First, the model generates a topic, z_n, associated with the region, using a multinomial distribution. Second, it generates a feature vecto,r r_n, conditioned on the topic, using a parameterized distribution, which is defined by the user. Third, it generates the local label associated with the region w_n, which also depends on the topic z_n, using a multinomial distribution. This local label also depends on the topic z_n. After performing the generative process N times, the image label is obtained by randomly selecting one of the regions and copying the local label.

The parameters of the distributions used in the three generative steps described above have to be estimated from the training data. A maximum likelihood estimation of the model parameters is unfeasible because it is not possible to analytically compute the likelihood function. The estimation is accomplished by resorting to variational methods, namely using a variational Expectation-Maximization algorithm (see [5] for details).

Given a new image, we apply the estimated corr-LDA model and assign the most probable label to each region. This provides a segmentation of the image according to the medical criteria. The next step concerns the estimation of a global label from the model information. We train a Random Forest classifier to predict the presence of each criteria. The inputs of the classifier are the probability distribution of each label given all the



Automatic labels: Dark brown black, and blue.

Automatic labels: Pigment network, dots, and blue-whitish veil.

Figure 2: Input image (1st row) and the output of the detection color (left) and structures (right) models (2nd row). The color scheme is the following. Right image: the color scheme is green for pigment network, red for dots/globules, and blue for blue whitish veil.

features extracted from the image, $p(w|\mathbf{r})$ [3].

After performing the detection of the medical criteria, our second goal is to use them to predict a lesion diagnosis. In this stage we follow a more traditional pattern recognition approach, *i.e.*, extract features from the detected criteria (outputs of corr-LDA) and use them to train a classifier. The extracted features are: i) the presence/absence of each criteria,ii) $p(w|\mathbf{r})$, and iii) the average number of regions per topic, computed as described in [5]. Different classification algorithms have been tested and the best one was Random Forests. Thus, this is the algorithm used in this work.

3 Experimental Results

The experiments were carried on a heterogeneous dataset of 804 images (241 melanomas) from the EDRA database [2]. All of the images were analyzed by several experts during a consensus meeting. Each image is associated with a set of global text labels stating which are the observed criteria. The training and test of the annotation and classification blocks were performed using a 10-fold nested cross-validation procedure.

Since color and structures co-occur, two corr-LDA models were trained, one for color and another one for structures. This allows the assignment of two different labels to the same region. In the case of the color model, the features r_n used to describe the regions are the mean HSV values, while in the case of the structures model the features are the mean HSV values and the texture features: contrast and anisotropy.

Figure 2 shows the output of the criteria detection block for both color and texture. The model is able to correctly predict all the color and structure labels associated with the image. It also provides a segmentation of the image according to the different criteria. Although we do not have ground-truth segmentation for the image (recall that the model was trained using text labels only), the segmentation proposed by the model seems to provide a correct interpretation of the image. Nonetheless, it would be interesting to be able to validate the segmentations comparing with the medical performance. However, this is far from being simple.

Tables 1 and 2 and show the performance of label assignment. The detection of each criteria is considered as a binary decision problem, characterized in terms of precision and recall. The system correctly detects most of the structures and the colors. However, it is possible to see that the performance changes according to the structure or color considered, e.g., dark brown is detected with a precision of 95.7% and recall 95.7%, while the red and white colors are detected with lowers scores. One possible explanation is the number of examples in the training set. Brown color is very common, while white and red are rare and appear only in 24 and 39 images, respectively.

Melanoma detection can be performed using color, structures or both. We trained a separate Random Forest for color and structures. Both of them can predict the presence of melanoma and give a score $s \in [0,1]$, where 1 is malignant and 0 is benign. We also combined their outputs using late fusion by simple average of both scores [7].

The performance of melanoma detection is shown in Table 3. Color and structures achieve comparable performances (structures perform slightly better). The combination of both criteria improves sensitivity, which is the most important metric (probability of correct detection if the lesion is a melanoma).

Table 1: Results for structure detection.

Structures	Precision	Recall
Pigment Network	78.5%	86.1%
Dots	72.8%	83.7%
Blue-whitish veil	82.8%	68.1%
Regression areas	63.9%	58.8%

Table 2: Results color detection.

Colors	Precision	Recall
Blue-Gray	87.6%	94.2%
Dark-Brown	95.7%	95.7%
Light-Brown	89.1%	92.7%
Black	81.5%	88.8%
Red	79.3%	74.2%
White	63.6%	93.3%

Table 3: Results for melanoma diagnosis using Random Forests.

Criteria	Sensitivity	Specificity
Structures	80.9%	74.8%
Colors	80.1%	71.9%
Combined	85.8%	71.1%

4 Conclusions

This paper proposes a system that extracts medical information from the image: it provides text labels of medical criteria and their location inside the lesion. This is achieved by training the system with weakly annotated images, which means that the training set is annotated by experts with text labels but no information is provided regarding their location within the image. Furthermore, the system uses this information to provide a diagnosis of the lesion as benign or melanoma. This means that dermatologists receive an automatic decision and the medical information that justifies it. To the best of our knowledge, this is the first system that provides this information.

References

[1] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and El. Delfino. 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. *Archives of Dermatology*, 134:1563–1570, 1998.

[2] 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.

[3] C. Barata, M. E. Celebi, J. S. Marques, and J. Rozeira. Clinically inspired analysis of dermoscopy images using a generative model. *accepted for publication in Computer Vision and Image Understanding*.

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

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

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2): 167–181, 2004.

[7] J. Kittler, M. Hatef, R.P. W. Duin, and J. Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

[8] K. Korotkov and R.l Garcia. Computerized analysis of pigmented skin lesions: a review. *Artificial intelligence in medicine*, 56(2):69–90, 2012.

[9] 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*, 4:521–527, 1994.