What is the Role of Annotations in the Detection of Dermoscopic Structures?

Bárbara Ferreira, Catarina Barata[0000−0002−2852−7723], and Jorge S. Marques[0000−0002−3800−7756]

Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal
ana.c.fidalgo.barata@tecnico.ulisboa.pt

Abstract. There has been an increasing demand for computer-aided diagnosis systems to become self-explainable. However, in fields such as dermoscopy image analysis this comes at the cost of asking physicians to annotate datasets in a detailed way, such that they simultaneously identify and manually segment regions of medical interest (dermoscopic criteria) in the images. The segmentations are then used to train an automatic detection system to reproduce the procedure. Unfortunately, providing manual segmentations is a cumbersome and time consuming task that may not be generalized to large amounts of data. Thus, this work aims to understand how much information is really needed for a system to learn to detect dermoscopic criteria. In particular, we will show that given sufficient data, it is possible to train a model to detect dermoscopic criteria solely using *global annotations* at the image level, and achieve similar performances to that of a fully supervised approach, where the model has access to *local annotations* at the pixel level (segmentations).

Keywords: Skin cancer · Dermoscopic structures · Supervised model · Weakly supervised model · corr-LDA.

1 Introduction

Annual reports, such as [13], show that the incidence rates of skin cancer, in particular of its most aggressive form (melanoma), have been steadily increasing for the past decades. Although dermoscopy has been shown to be a powerful imaging technique, the diagnosis of dermoscopy images remains subjective a challenge for untrained dermatologists [2]. This has favored the development of computer-aided diagnosis (CAD) systems that can perform a preliminary screening of the dermoscopy images and work as second opinion tool, from which inexperienced doctors may learn [11, 12].

One of the most challenging aspects of CAD systems is the selection of discriminative features to characterize the dermoscopy images [4]. Earlier works based their image descriptors on the medical ABCD rule, which amounted to compute abstract asymmetry, border, color, and texture descriptors, while most recent methods avoid the process of feature design and rely on deep learning architectures to learn the most discriminative features. In both cases, the features lack medical meaning, making it hard to: i) understand the diagnosis and

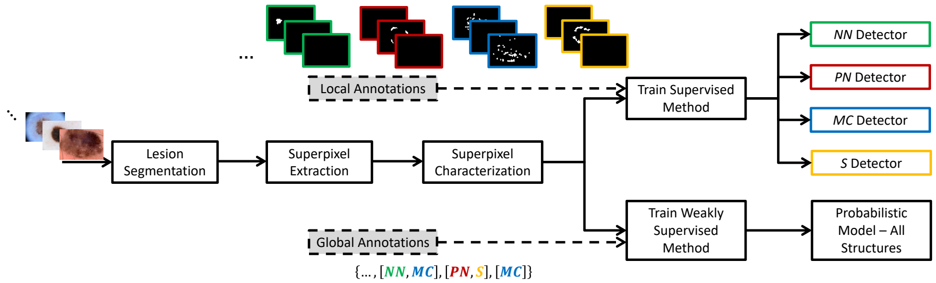

Fig. 1. Block diagram of the training of the proposed models. The colors represent the different dermoscopic structures: NN - negative network (green); PN - pigment network (red); MC - milia-like cysts (blue); and S - streaks (orange).

ii) be accepted by the medical community. This issue has fostered a promising line of research, where the goal is to develop clinically oriented systems that extract features that have a medical meaning, namely dermoscopic structures and colors [8]. However, this is a challenging field, due to the subtlety of several of the dermoscopic criteria and the need for extensively annotated datasets. These datasets must not only comprise the diagnosis of a skin lesion, but also detailed information of the observed dermoscopic criteria, such as segmentations at the pixel level, which we will refer to as *local annotations*. Until very recently, a dataset of this kind was not publicly available, which fomented the proposal of weakly supervised methods [3, 10] that could deal with the only available information (*global annotations*, *i.e.*, labels at the image level), and still be able to localize the criteria in the skin lesions. Fig. 1, shows the difference between *local* and *global annotations*.

Very recently, the International Skin Imaging Collaboration (ISIC), released a dataset that contained more than 2000 images, each with *local annotations* for a small set of relevant dermoscopic structures (pigment network, streaks, milia-like cysts, and negative network) [9]. This dataset has the potential to push forward the development of clinically inspired systems. Nonetheless, several dermoscopic criteria are still missing and the number of examples for most of the structures is small.

For dermatologists to provide *local annotations* is a cumbersome and time consuming task. Thus, it is important to understand if this information is really needed or if it is possible to develop a clinically inspired method, solely using *global labels*. This paper addresses the aforementioned issue and performs a comparison between a fully supervised system, which learns to detect dermoscopic criteria from *local annotations*, and a weakly-supervised system that only uses *global annotations* to perform the same task.

2 Proposed Framework

The goal of this work is to compare the performance of a supervised learning system against a weakly supervised one on the task of localizing four differ-

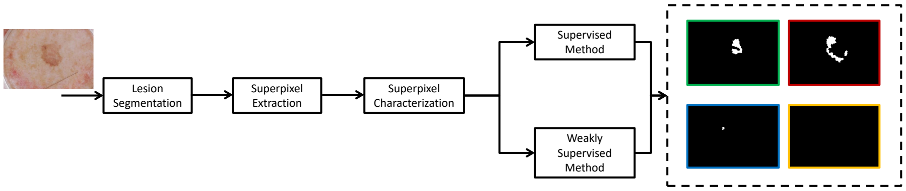

Fig. 2. Block diagram for testing the proposed models and desired output. The colors-lined masks represent the localization of the dermoscopic structures: negative network (green); pigment network (red); milia-like cysts (blue); streaks (orange).

ent dermoscopic structures (pigment network, negative network, streaks, and milia-like cysts) in skin lesions. The block diagram for training and testing the methods is shown in Figs. 1 and 2. The first diagram clearly states the main different between the supervised and weakly supervised strategies: the first relies on labels at the pixel level, called *local annotations*, thus has access to very detailed information, while the latter relies on less informative text labels at the image level (*global annotations*). The second diagram shows that we expect to obtain the same output with both methods, *i.e.*, a set of binary masks with the location of the structures.

Below, we succinctly describe the steps that are common to the two methods. The supervised method is detailed in Section 3, while the weakly supervised approach is described in Section 4.

2.1 Lesion Segmentation and Superpixel Extraction

Skin lesions are segmented in order to separate them from the surrounding healthy skin. In this work we use manual segmentation masks performed by experts [9].

The following step is to divide the lesion into small and homogeneous regions, called superpixels, such that each region can be analyzed independently and classified regarding the presence/absence of each dermoscopic structure. The algorithm used to compute the superpixels is SLIC0 [1, 9].

2.2 Superpixel Characterization

The superpixel algorithm leads to the segmentation of the skin lesion into N regions that must be characterized by a feature vector $r_n \in \mathbb{R}^f$, such that an image d may be represented by the following set $\mathbf{r}^d = \{r_1^d, ..., r_N^d\} \in \mathbb{R}^{f \times N^d}$, where N^d is the number of superpixels of that image. We rely on color and texture information to characterize the superpixels, in particular:

Color features: This property is characterized using the mean color vector in the HSV space (μ_{HSV}).

Texture features: Three types of descriptors are used to characterize the texture of a region. The mean contrast (μ_{c}), the mean contrast × anisotropy (μ_{ca}), and statistics computed using the directional filters from [5]. Both contrast and anisotropy are computed from the second moment matrix estimated at each pixel $M(x,y)$

$$a(x,y) = 1 - \tfrac{\lambda_2}{\lambda_1},\ c(x,y) = 2\sqrt{\lambda_1 + \lambda_2}\,, \tag{1}$$

where λ_1, λ_2 are the eigenvalues of $M(x,y)$ [7]. The directional filters are applied at different orientations $\theta_i \in [0,\pi]$, $i = 0,\dots,9$, with the impulse response for any direction θ_i given by

$$h_{\theta_i}(x,y) = G_1(x,y) - G_2(x,y)\,, \tag{2}$$

where G_k is a Gaussian filter:

$$G_k(x,y) = C_k \exp\{-\frac{x'^2}{2\sigma^2_{x_k}} - \frac{y'^2}{2\sigma^2_{y_k}}\}, k = 1,2\,. \tag{3}$$

C_k is a normalization constant and the values of (x',y') are related with (x,y) by a rotation of amplitude θ_i.

$$\begin{aligned} x' &= x\cos\theta_i + y\sin\theta_i\,,\\ y' &= y\cos\theta_i - x\sin\theta_i\,. \end{aligned} \tag{4}$$

We compute the output of the directional filters (2) for all the directions and keep the maximum and minimum at each pixel (x,y). The regions are described by the mean and standard deviation of these values (μ_M, σ_M, μ_m, and σ_m).

3 Supervised Model

In the supervised context, the model has access to all detailed information during the training phase. In the case of this work, this means that to be able to train a supervised model, we need to have access to detailed ground-truth annotations at the pixel or superpixel level (*local annotations*, as shown in Fig.3). The annotations used in this work are at the superpixel level.

Given the *local annotations*, the problem of localizing dermoscopic structures becomes a simple classification problem, where our goal is to classify each superpixel into one of four possible classes: negative network, pigment network, milia-like cysts, and streaks. Although this could be treated as a multi-class problem, some of the superpixels have more than one label (see Fig.3). Thus, we will train a separate classifier for each of the classes, as shown in Fig.1. During the test phase, each of the classifiers is separately used to label the superpixels.

The classification algorithm used in this work is SVM (support vector machines) with an radial basis function (RBF) kernel.

4 Weakly Supervised Model

The supervised approached described in Section 3 corresponds to the ideal scenario, where we have access to detailed annotations of the presence and location

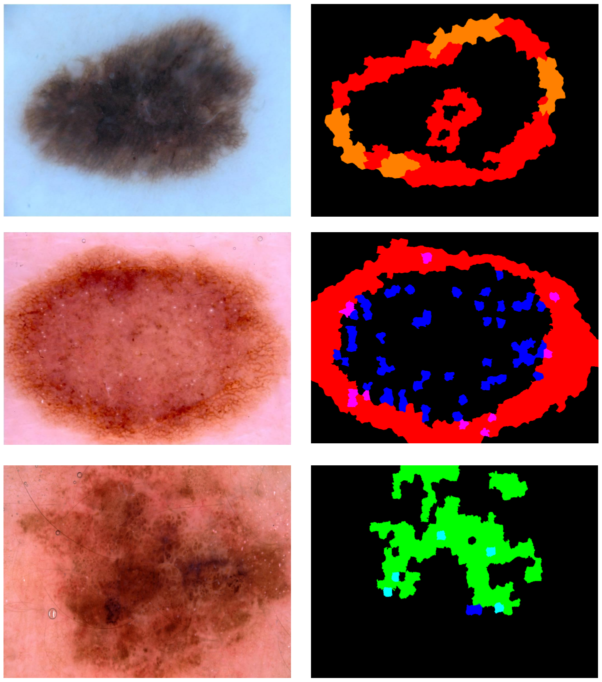

Fig. 3. *Local annotations*: negative network (green); pigment network (red); milia-like cysts (blue); and streaks (orange). The remaining colors identify superpixels that have multiple labels.

of the dermoscopic structures that we want to detected. However, until the recent release of challenge related datasets, such as ISIC 2017 [9], this kind of information was unavailable. Most of times, one has only access to the dermoscopy images and a set of *global annotations* at the image level, swhich dermoscopic structures are present. To address this limitation and still be able to localize the dermoscopic structures, Barata et al. [3] proposed a framework based on the correspondence latent Dirichlet allocation (corr-LDA) model [6].

Corr-LDA belongs to the family of generative algorithms for image captioning [6]. The main idea of this method is that we can represent an image as a distribution over a set of K latent variables z, called topics. Each of the topics allow us to simultaneously express: i) a distribution over superpixel features $p(r_n|z_k, \Omega_k)$, where Ω_k is the set of parameters of the distribution associate to z_k; and ii) a multinomial distribution over the possible *global annotations* $p(w|z_k, \beta_k)$ with parameter β_k. Thus, through the topics, we are able to find the relationship between superpixel features r_n and the *global annotations*, and consequently label each superpixel.

Training the corr-LDA model amounts to estimating the set of parameters $\{\Omega_1, \ldots, \Omega_K, \beta_1, \ldots, \beta_K\}$, given a set of superpixel features for different images $\mathcal{R} = \{\mathbf{r}^1, \ldots, \mathbf{r}^D\}$ and their corresponding *global annotations* (negative network, pigment network, milia-like cysts, and streaks). We use the strategy described

Table 1. ISIC 2017 statistics: PN - pigment network; MC - milia-like cysts; NN - negative network; and S - streaks.

	Local Annotations					*Global Annotations*				
Set	**# SPixel**	**% PN**	**% MC**	**% NN**	**% S**	**# Images**	**% PN**	**% MC**	**% NN**	**% S**
Train	460272	16.92	1.01	0.71	0.46	2000	56.40	28.40	6.25	5.80
Val.	31946	10.41	1.02	1.03	0.04	150	42.67	28.67	7.33	1.33
Test	193730	10.38	0.66	1.12	0.07	600	55.50	24.50	7.50	1.50

in [6,3], to train our model. On the test phase, we apply the estimated model and compute the probability $p(w|r_n)$ (see [3]), to determine the probability of each global label. Since some of the superpixels may be associated with more than one label (recall Fig.3), we experimentally determined a threshold on the probability of each annotation, such that a dermoscopic structure was assigned to a superpixel if its probability was greater than the threshold.

5 Experimental Results

5.1 Dataset and Performance Metrics

The supervised and weakly supervised approaches were evaluated using the ISIC 2017 challenge dataset [9], which comprises 2750 images divided into training, validation, and test sets. The training set was used to estimate parameters of the supervised (SVM) and weakly supervised (corr-LDA) models, while the validation set was used to selected the best model and respective hyperparameters (the kernel width in RBF and the number of topics K in corr-LDA). The test set was used to evaluate and compare the models.

Each of the images was segmented by experts, who also annotated each of the superpixels into one or more of the following dermoscopic structures: negative network, pigment network, milia-like cysts, and streaks. Table 1 shows the number of superpixels per set, as well as the percentage that is associated to each structure. Since the weakly supervised model uses image-level annotations, we have defined that an image receives the *global annotation* of a specific structure if it has at least one superpixel with that label. For computational reasons it was necessary to define an additional *global annotation* called "without structure" to deal with images that do not exhibit any structure. The proportion of *global labels* per type of structure is also shown in Table 1.

These statistics show that there is a significant imbalance in the dataset, with pigment network being the most common annotation both at the superpixel and image level. To overcome this limitation, we used two strategies: i) in the supervised approach, we have assigned different weights to the classes during the training of the classifiers, based on their distribution; ii) in the weakly supervised approach we have artificially augmented the data, such that there were at last 500 images for each type of *global annotation.*

The metrics used to evaluate the models are the sensitivity (SE), specificity (SP), and balanced accuracy ($BACC$)

$$SE = \frac{TP}{TP+FN}\,,\ SP = \frac{TN}{TN+FP}\,,\ BACC = \frac{SE+SP}{2}\,, \tag{5}$$

Table 2. Experimental results for the supervised and weakly supervised models. In **bold** we highlight the most interesting scores.

	Supervised			**Weakly Supervised**		
Dermoscopic Structure	SE	SP	$BACC$	SE	SP	$BACC$
Pigment Network	**84.6%**	**69.2%**	**76.9%**	**73.3%**	**76.0%**	**74.7%**
Milia-like Cysts	62.6%	60.3%	61.5%	0.6%	98.6%	49.6%
Negative Network	67.6%	70.8%	69.2%	3.4%	99.3%	51.3%
Streaks	71.4%	73.0%	72.2%	25.7%	94.5%	60.1%

where TN, TP, FN, and FP are the total number of superpixels that are respectively true negatives, true positives, false negatives, and false positives.

5.2 Results

The scores for the supervised and weakly supervised models are shown in Table 2. In the case of the supervised model, the classification of the superpixels w.r.t to each dermoscopic structure is achieved with different degrees of success. In particular, it seems that milia-like cysts and negative network are harder to identify than the other two structures.

The weakly supervised model exhibits low SE for milia-like cysts, negative network, and streaks. However, the most noteworthy result is that of pigment network. The performance for this structure is similar to that of the supervised approach at the cost of approximately 0.43% for the annotations: recall that the supervised method relies on more than 460K *local annotations* at the superpixel level, while the weakly supervised uses 2K *global annotations* at the image level. This suggests that given a sufficient number of images with a specific dermoscopic structure, one does not require very detailed annotations as a weakly supervised model seems to be able to achieve a similar performance. The streaks performance seems to also support this claim, as with only 5.8% of the images with this annotation, corr-LDA still achieves a $SE = 25.7\%$ and a $SP = 94.5\%$. Since providing detailed annotations is a cumbersome task, this result opens new possibilities for training models with less information.

Negative network and especially milia-like cysts seem to be the structures that achieve the worse performances in both supervised and weakly-supervised approaches. This suggests that the features used to characterize the superpixels may not efficiently represent these two structures. Thus, a future direction of improvement will be to replace the features described in Section 2.2 with more powerful descriptors, namely those based on deep neural networks.

6 Conclusions

This paper compares two models for the detection and localization of four dermoscopic structures (pigment network, milia-like cysts, negative network, and streaks) in superpixels of dermoscopy images. The first model was fully supervised, thus was trained using more than 4.60×10^5 *local annotations* at the superpixel level, while the second model relied on a weakly-supervised algorithm (corr-LDA), trained using 2×10^3 *global annotations* at the image level. The experimental results surprisingly showed that in the case of pigment network, the weakly supervised model is able to achieve a similar performance to that of the supervised one, with a significantly smaller amount of ground truth information. However, the supervised model achieved better performances for the remaining structures, suggesting that better features and more data is required. Nonetheless, the most relevant output of this work is that given a sufficient number of images, we may not need detailed information to be able to detect dermoscopic structures.

7 Acknowledgments

This work was supported by the FCT project and plurianual funding: [PTDC/EEI PRO/0426/2014], [UID/EEA/50009/2019].

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
2. Argenziano, G., Soyer, H.P., De Giorgi, V., et al.: Interactive Atlas of Dermoscopy. EDRA Medical Publishing & New Media (2000)
3. Barata, C., Celebi, M.E., Marques, J.S.: Development of a clinically oriented system for melanoma diagnosis. Pattern Recognition **69**, 270–285 (2017)
4. Barata, C., Celebi, M.E., Marques, J.S.: A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE Journal of Biomedical and Health Informatics (2018)
5. Barata, C., Marques, J.S., Rozeira, J.: A system for the detection of pigment network in dermoscopy images using directional filters. IEEE Transactions on Biomedical Engineering **59**(10), 2744–2754 (2012)
6. Blei, D., Jordan, M.: Modeling annotated data. In: 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 127–134. ACM (2003)
7. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(8), 1026–1038 (2002)
8. Celebi, M.E., Codella, N., Halpern, A.: Dermoscopy image analysis: Overview and future directions. IEEE journal of biomedical and health informatics (2019)
9. Codella, N.C.F., Gutman, D., Celebi, M.E., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172 (2018)
10. Madooei, A., Drew, M.S., Hajimirsadeghi, H.: Learning to detect blue-white structures in dermoscopy images with weak supervision. IEEE Journal of Biomedical and Health Informatics (2018)

11. Oliveira, R., Papa, J., Pereira, A., Tavares, J.: Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Computing and Applications pp. 1–24 (2016)
12. Pathan, S., Prabhu, K.G., S., P.C.: Techniques and algorithms for computer aided diagnosis of pigmented skin lesions - a review. Biomedical Signal Processing and Control **39**, 237–262 (2018)
13. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: a cancer journal for clinicians **69**, 7–34 (2019)