31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Variability in human and automatic segmentation of melanocytic lesions

A. Silletti, E. Peserico, A. Mantovan, E. Zattra, A. Peserico, A. Belloni Fortina

Abstract— **In a double blind evaluation of 60 digital dermatoscopic images by 4 "junior", 4 "senior" and 4 "expert" dermatologists (dermatoscopy training respectively less than 1 year, between 1 and 5 years, and more than 5 years), a significant inter-operator variability was observed in melanocytic lesion border identification (with a disagreement of the order of $10-20\%$ of the area of the lesions). Expert dermatologists showed greater agreement among themselves than with senior and junior dermatologists, and a slight tendency towards "tighter" segmentations.**

The human inter-operator variability was then used to evaluate the segmentation accuracy of 4 algorithms, representative of the 3 fundamental state-of-the-art automated segmentation techniques and of a fourth, novel, technique. Our evaluation methodology addresses a number of crucial difficulties encountered in previous studies and may be of independent interest. 3 of the 4 algorithms showed considerably less agreement with expert dermatologists than even senior and junior dermatologists did (with a disagreement of the order of 30% of the area of the lesions); the remaining algorithm, however, showed agreement with expert dermatologists comparable to that of other expert dermatologists.

I. Introduction

The first step in the analysis of any dermatoscopic image of a melanocytic lesion is *segmentation*, i.e. classification of all points in the image as part of the lesion or simply part of the surrounding, healthy skin. While segmentation is typically studied in the context of automated analysis of images, it is important to observe that it is a first, necessary step even for human operators who plan to evaluate quantitative features of a lesion such as diameter or asymmetry - e.g. in the context of epidemiological studies correlating those features to lesion benignity [12].

Unfortunately, segmentation of melanocytic lesions is a surprisingly difficult task, for human operators and automated systems alike. The fundamental reason lies in the fact that lesion borders are often fuzzy and there exists no standard operative definition of whether a portion of skin belongs to a lesion or not. Dermatologists rely on subjective judgement developed over years of dermatoscopic training. Automated systems attempt to replicate the assessment of human dermatologists through a number of heuristics. Not surprisingly, this leads to appreciable variability in the localization of the precise border of lesions, not only between automated systems and human dermatologists, but also between different human dermatologists [9].

Quantifying this variability is crucial for at least two reasons. First, it allows one to estimate the level of noise affecting large, multi-operator epidemiological studies e.g. correlating lesion size to benignity. Second, human inter-operator variability effectively provides an upper bound to the segmentation accuracy achievable by any automated system, as long as "ground truth" is provided by the subjective evaluation of human dermatologists rather than by a standard operative definition. For example, if even experienced dermatologists disagree on how to classify 5% of the area of an image, no automated system can be expected to classify "correctly" more than 95% of the area of that image.

This paper evaluates the variability in lesion border identification by a group of 12 dermatologists. This is the largest such study so far, and the only one that differentiates dermatologists based on dermatoscopy training experience; our evaluation methodology also addresses some crucial difficulties inherent to previous studies. The human inter-operator variability is then compared to the segmentation accuracy of four algorithms, representative of the three fundamental state-of-the-art automated segmentation techniques and of a fourth, novel, technique.

The rest of the paper is organized as follows. Section II reviews the current metrics used to evaluate inter-operator variability and automated segmentation accuracy, discussing some difficulties inherent the most sophisticated approaches – and how to address them. Section III describes the details of the segmentation experiment involving 60 dermatoscopic images, and the results in terms of (human) inter-operator variability; it also reviews the fundamental techniques for melanocytic lesion segmentation and evaluates an algorithm representative of each within the framework introduced in Section II. Section IV summarizes our results and discusses their significance before concluding with the bibliography.

II. Measuring variability and accuracy

We shall see that the two issues of measuring inter-operator variability in segmentation of melanocytic lesions and of measuring accuracy of automated segmentation methods are closely related. The scant body of work on the former (essentially amounting to [9]) seems then surprising given the vast literature dealing with the latter (e.g. [1], [11], [16], [7], [4], [6]) that we briefly review below.

While some studies (e.g. [16], [6]) have one or more dermatologists subjectively assess the quality of the proposed automated systems, the general consensus is that evaluation methods striving for a greater degree of objectivity are preferable [4]. Most of these methods rely on a *ground truth* segmentation against which the proposed segmentation is assessed, labeling its pixels as True Positive (TP), False Positive (FP), False Negative (FN) or True Negative (TN),

Supported in part by the University of Padova "Naevi in silico" project (melanoma@dei.unipd.it). The first three authors are affiliated with the Dep. of Information Engineering and the last three with the Dermatology Unit.

TABLE I

SOME COMMON METRICS TO EVALUATE SEGMENTATION

XOR Error [2]	$(\frac{FP+FN}{TP+FN}) \times 100\%$
Specificity	$(1 - \frac{FP}{FP+TN}) \times 100\%$
Sensitivity or Recall	$(1 - \frac{FN}{TP+FN}) \times 100\%$
Precision	$(1 - \frac{FP}{FP+TP}) \times 100\%$

depending on whether they are classified as part of the lesion, respectively, in both segmentations, only in the proposed segmentation, only in the ground truth segmentation, or in neither of the two. The number of pixels in the FP and FN categories, usually normalized dividing them either by the size of the proposed lesion (TP+FP), by the size of the ground truth lesion (TP+FN) or by the size of its complement (FP+TN), provide a measure of the divergence between the proposed segmentation and the ground truth.

The fundamental problem with these approaches is that any definition of "ground truth" based on the segmentation of a single dermatologist is inherently highly subjective. Thus, several recent approaches combine the evaluation of multiple dermatologists to obtain a more objective ground truth segmentation. However, these approaches are more complex, and all exhibit some shortcomings.

[1] compares the proposed segmentation with the segmentation of each "ground truth" dermatologist separately without attempting a summary – this makes it hard to compare two different proposed segmentations. [10] obtains a single ground truth segmentation from those of multiple dermatologists through simple majority voting – as observed in [4] this does not discriminate between a situation with high consensus and one with high divergence between different ground truths, whereas the former should intuitively penalize a divergence of the proposed segmentation more heavily than the latter. [4] proposes the use of the NPRI metric, first introduced in [19] to assess the quality of generic segmentation algorithms. Unfortunately, NPRI is very complex, lacking the immediacy of True/False Positive/Negative approaches, and exhibits some highly counterintuitive behaviors [14]: e.g. a segmentation that is "tighter" than ground truth may receive a *worse* score than a segmentation that is even "tighter" (see Figure 1). [9] computes, for each proposed segmentation, a *Misclassification probability* that is essentially the average (over all ground truths) of $\frac{FP}{FP+TP}$ – i.e. the average fraction of the segmented lesion misclassified as lesion. Unfortunately, this does not penalize false negatives at all, and all segmentations "tighter" than ground truth receive the same score as ground truth itself.

In addition to their individual shortcomings, all these metrics share a subtler, but perhaps more serious problem: they do not provide an idea of how well one can expect a proposed segmentation to perform. While some of them (e.g. [10], [4]) are normalized in such a way that, for every (set of) ground truth(s), the best score a segmentation can achieve is exactly 1, it is generally unrealistic for any human

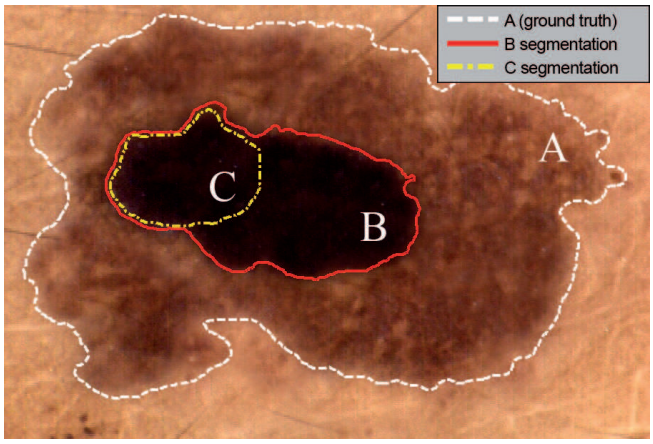

Fig. 1. Counterintuitive behavior of NPRI and Misclassification Probability. The latter assigns the same score to the the ground truth segmentation (A) and to the two "tighter" segmentations (B and C). NPRI assigns to B a score worse than C, despite the fact that B is closer to the ground truth.

dermatologist – and thus for any automated segmentation system – to achieve such a score.

The solution we propose is simple, and in the spirit of the classic Turing test [17]: when evaluating an automated segmentation system, in addition to the ground truth segmentation(s), one should always employ one more "calibration" segmentation provided by an experienced dermatologist. The divergence of the calibration segmentation from the ground truth (by whatever metric one may choose) provides a clear, intuitive indication of the best divergence one can hope for when evaluating by that same metric an automated segmentation system (or even, in fact, a less experienced dermatologist!). *In a nutshell, we propose the variability between experienced human dermatologists in the localization of melanocytic lesion border to be used as a* gold standard *to assess the quality of any automated segmentation system.*

While the choice of the basic divergence metric is relatively unimportant, our choice would fall on the average of $\frac{FP}{FP+TP}$ over all ground truths (i.e. the Misclassification probability of [9]) paired with the complementary average of $\frac{FN}{FP+TP}$ to account for false negatives. The latter metric is similar to Precision and Recall, but the normalization takes place over the size of the lesion according to the segmentation under test (as in [9]) rather than according to ground truth – allowing divergence from each ground truth to have the same weight. This pair of metrics makes extremely clear the source of a segmentation's divergence from ground truth - identifying whether the cause lies in many ground truth lesion pixels classified as healthy skin (leading to high FN) or many ground truth healthy skin pixels classified as lesion (leading to high FP).

III. COMPARING DERMATOLOGISTS AND ALGORITHMS

The (768 by 576 pixel) images of 60 melanocytic lesions were acquired using a Fotofinder digital dermatoscope. 12 copies of each image where then printed on 13 cm by 18 cm photographic paper. A copy of each image together with a marker was given to each of 4 "junior", 4 "senior" and

4 "expert" dermatologists (respectively less that 1 year of dermoscopy training, between 1 and 5 years, and more than 5 years). Each dermatologist was then asked to independently draw the border of each lesion with the marker. The images (and borders) were scanned and realigned to the same frame of reference. Finally, the contours provided by the markers were extracted and compared. This allowed the identification, for each pixel of each original image, of the set of dermatologists classifying it as part of the lesion proper or of the surrounding, healthy skin.

This approach required a considerable amount of engineering effort compared to that of similar studies in the literature. [9] had dermatologists use Adobe PhotoShop's "pencil" tool to draw a polygonal approximation of the contour. [4] and [2] had dermatologists indentify a sparse set of points in the contour and then fit the points to a second-order B-spline. [10] had dermatologists draw the border on a tablet computer. Our goal was to maximize the comfort of dermatologists, thus minimizing the noise in border localization caused by the use of unfamiliar drawing tools.

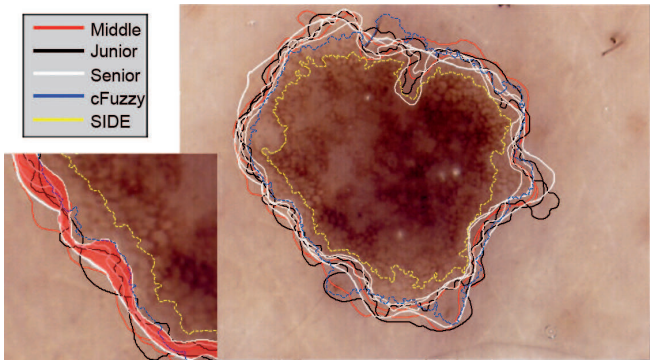

Fig. 2. Hand traced borders and two borders obtained with modified c-Fuzzy and SIDE

Each of the 4 possible sets of 3 expert dermatologists was used to provide a "ground truth" from which the divergence of the remaining expert dermatologist, of the 4 senior and the 4 junior dermatologists, as well as of 4 segmentation algorithms was assessed. Figure 3 shows the average value (over the 60 images and the 3 ground truth segmentations) of the values of $\frac{FP}{TP+FP}$ and of $\frac{FN}{TP+FP}$ (see Section II).

The 4 algorithms are representative of the 3 main classes of automated lesion segmentation techniques in the literature, as well as of a fourth, novel technique.

The first class uses edges and smoothness constraints to identify the lesion. We implemented GVF Snakes [7]: a promising approach, though with a number of serious shortcomings. The algorithm requires a good initial segmentation to converge, a preprocessing such as black frame removal or hair removal [3], [20], and a morphological postprocessing to refine the results.

The second class performs color clustering directly on the image: this includes Modified JSEG [2] and SIDE [8]. We implemented the latter.

The third class performs clustering on the color histogram and then maps back to the original image. Mean-

Ground Truth, average of:

	Experts 2, 3, 4		**Experts** 1, 3, 4		**Experts** 1, 2, 4		**Experts** 1, 2, 3	
	$\frac{FN}{TP+FP}$	$\frac{FP}{TP+FP}$	$\frac{FN}{TP+FP}$	$\frac{FP}{TP+FP}$	$\frac{FN}{TP+FP}$	$\frac{FP}{TP+FP}$	$\frac{FN}{TP+FP}$	$\frac{FP}{TP+FP}$
Cal.Expert	2.51%	11.10%	10.18%	3.07%	9.10%	3.56%	5.21%	5.70%
Senior 1	1.50%	10.58%	3.15%	8.41%	3.06%	8.74%	2.85%	9.72%
Senior 2	1.60%	13.23%	2.66%	10.71%	2.65%	11.04%	2.39%	12.00%
Senior 3	5.19%	7.47%	7.29%	5.64%	7.01%	5.80%	6.67%	6.68%
Senior 4	6.27%	6.61%	8.84%	5.07%	8.58%	5.26%	8.10%	6.06%
Junior 1	1.96%	14.37%	3.15%	11.90%	3.05%	12.23%	2.91%	13.25%
Junior 2	0.85%	14.51%	2.19%	12.17%	2.15%	12.52%	1.99%	13.53%
Junior 3	1.01%	17.63%	1.80%	14.96%	1.87%	15.32%	1.68%	16.34%
Junior 4	1.16%	11.02%	3.22%	8.88%	3.14%	9.21%	2.94%	10.19%
c-Fuzzy	5.34%	5.62%	8.21%	4.28%	8.11%	4.69%	7.32%	5.28%
SIDE	23.58%	2.88%	28.46%	2.12%	28.33%	2.31%	26.83%	2.69%
Stat. Thre.	26.39%	11.29%	24.84%	11.17%	25.23%	12.31%	22.80%	12.23%
Snakes	10.03%	19.36%	12.27%	17.66%	12.47%	18.02%	11.46%	18.62%

Fig. 3. Average divergence of each expert dermatologist from the ground truth provided by the other three; and average divergence of senior and junior dermatologists and of 4 segmentation algorithms from the same ground truth. Divergence is measured as false negative area (FN: lesion pixels misclassified as healthy skin) and false positive area (FP: healthy skin pixels misclassified as lesion) as a percentage of the proposed segmentation area (TP+FP: pixels corretly or incorrectly classified as lesion).

Shift [11] and Fuzzy c-means [16] are representative of this class. These clustering algorithms work either using the RGB space [1], the B component [10], the Lab space [21], or the Pricipal Component decomposition [16]. We implemented an algorithm close in spirit to Fuzzy c-means – and that overcomes the initialization problem. The algorithm computes the Principal Components of the image, using the Karhunen-Love transform, and then the 2-dimensional histogram h associated with the two components of largest variance. Given the number of clusters, the two following two (recursive) equations optimally cluster the color space:

$$U_{k,\mathbf{x}} = \frac{1}{d(\mathbf{x},\mathbf{c}_k)^\alpha}$$

$$\forall k: \quad \arg\min_{\mathbf{c}_k} \quad I_k = \int_A (U_{k,\mathbf{x}})^\beta |\mathbf{x},\mathbf{c}_k|^\gamma h(\mathbf{x})^\lambda d\mathbf{x}$$

where $\mathbf{x}$ is a point in the 2d color space, $\mathbf{c}_k$ is the center of cluster k in the color space, $U_{k,\mathbf{x}}$ refers to the fuzzy membership of $\mathbf{x}$ to cluster k, $|\cdot|$ is the Euclidean distance, α,β,γ and λ are scalar values and A is the image. A steepest descent algorithm performs the minimization until a steady state is reached. The histogram clustering is then mapped back to the original image and a morphological postprocess removes the smallest areas.

A technique that does not fit into any of the three classes above could be based on *Statistical Thresholding* - in a nutshell, classifying as lesion those portions of skin that statistically differ in color from healthy skin. Given the average RGB color μ and matrix variance Σ of an healthy patch of skin (e.g. taken from the boundaries of the image) each pixel is classifies as lesion according to

$$d(c,\mu) \geq k \cdot |\Sigma|$$

where d is the Euclidean distance in the color space and k is a scalar controlling the sensitivity of the algorithm. Obviously, the algorithm does not perform well on lesions covering only a small region of the image: this is a problem

common to many algorithms that can be easily fixed with a crop of the image frame. The advantages of this approach are that it is simple to implement, and that it corresponds to a very "natural" definition of lesion (as the portion of skin exhibiting sufficient color variance from healthy skin).

IV. Discussion and conclusions

The results of Figure 3 show appreciable variability in the localization of the border of melanocytic lesions between human dermatologists. Even an expert dermatologist "misclassifies" (compared to a ground truth provided by other expert dermatologists) a portion of the image with an area between 2.2% and 39.1% of the area of the lesion itself. Less experienced dermatologists have an even lower agreement with their expert colleagues: the misclassified portion of the image has an area between 7.4% and 62.5% of the area of the lesion itself for "senior" and between 5.9% and 152.4% for "junior" dermatologists.

Although not entirely apparent from Figure 3, this divergence is not due to a systematic bias of individual dermatologists towards "tighter" or "looser" borders: we ranked all dermatologists for each lesion in order of increasing surface classified as lesion, and each dermatologist ranked first on at least one lesion, and at eighth or "larger" on at least another. On the other hand, expert dermatologists do show a very slight bias towards "tighter" borders (perhaps a symptom of greater confidence), and also, as should be expected, a somewhat greater agreement with other expert dermatologists than with less experienced ones.

It would certainly be interesting to study the impact of such variability on large, multi-operator epidemiological studies. These results seem to roughly confirm those of [9], though they are not directly comparable due to the different methodology ([9] evaluates the segmentation divergence of human dermatologists from a mix of human and algorithmic segmentations, rather than only from human segmentations). They also suggest that dermatoscopy skills require at least several years of training to mature.

In terms of algorithms, SIDE, Snakes and Statistical Thresholding did not perform very well, misclassifying a portion of the image with an area respectively between 8.4% and 92.6%, between 12.1% and 245.5%, and between 13.8% and 151.9% of the area of the lesion itself. These 3 algorithms were outperformed on average by every dermatologist, including ones belonging to the least experienced, "junior" cohort. As for Snakes, this might have been expected, in the light of the recent results of [5]. As for Statistical Thresholding, this shows that unfortunately the most natural, axiomatic definition of lesion (as the portion of skin exhibiting sufficient color variance from healthy skin) fails to provide results that, in practice, match the actual intuition of the human eye. As for SIDE, its poor performance is somewhat unexpected, given the results of [8]. This may be due, in part, to the fact that SIDE is a particularly difficult algorithm to calibrate correctly - its performance could perhaps be improved with better fine-tuning than what we managed to achieve.

On the other hand, our variant of Fuzzy c-means performed extremely well. On average, it misclassified a portion of the image with an area between 3.7% and 50.2% of the area of the lesion itself (again, using as ground truths the segmentations provided by teams of three expert dermatologists). This is barely worse, and in 1 case out of 4 better, than the performance of the fourth, expert dermatologist used as "control" in each case. It is also significantly better than the performance of all remaining senior and junior dermatologists. Figure 2 provides a visual intuition of the quality of the results of this algorithm. Fuzzy c-means thus appears an excellent candidate to provide standardized, objective and highly reproducible segmentation of melanocytic lesions and assessment of corresponding features that closely match those of the most experienced dermatologists.

References

[1] M. E. Celebi at al., Border Detection in Dermoscopy Images using Statistical Region Merging, *Skin Res. Tech.*, 14:347-353, 2008

[2] M. E. Celebi et al., Unsupervised Border Detection in Dermoscopy Images, *Skin Res. Tech.*, 13(4):454-462, 2007

[3] M. E. Celebi et al., Border Detection in Dermoscopy Images Using Statistical Region Merging, *Skin Res. Tech.* 14:347-353, 2008

[4] M. E. Celebi, G. Schaefer and H. Iyatomi, Objective evaluation of methods for border detection in dermoscopy images, *Proc. of IEEE EMBS*, 3056-9, 2008

[5] A. Cenedese and A. Silletti, A robust generalized Active Contour approach for studying cell deformation from noisy images, *ICMBE*, 2009

[6] R. Cucchiara et al., Exploiting Color and Topological Features for Region Segmentation with Recursive Fuzzy C-Means, *Mach. Graph. Vis.*, 11(2/3):169-182, 2002

[7] B. Erkol et al., Automatic Lesion Boundary Detection in Dermoscopy Images using Gradient Vector Flow Snakes, *Skin Res. Tech.*, 11:17-26, 2005

[8] J. Gao et al., Segmentation of Dermatoscopic Images by Stabilized Inversediffusion Equations, *Proc. ICIP*, 3:823-827, 1998

[9] J. Guillod et al., Validation of Segmentation Techniques for Digital Dermoscopy, *Skin Res. Tech.*, 8(4):240-249, 2002.

[10] H. Iyatomi et al., Quantitative Assessment of Tumour Extraction from Dermoscopy Images and Evaluation of Computer-Based Extraction Methods for an Automatic Melanoma Diagnostic System, *Melanoma Res.* 16(2):17-26, 2005

[11] R. Melli, C. Grana, R. Cucchiara, Comparison of Color Clustering Algorithms for Segmentation of Dermatological Images, *Proc. of SPIE* 6114, 2006

[12] F. Nachbar et al. The ABCD rule of dermatoscopy. *J. Am. Acad. Dermatol.*, 30:551-559, 1994

[13] R. Nock and F. Nielsen, Statistical Region Merging. *IEEE Trans. on Pattern An. and Machine Intell.*, 1452-1458, 2004

[14] E. Peserico and A. Silletti, Evaluating segmentation of melanocytic lesions with (N)PRI: an anomaly, *Univ. of Padova Tech. Report*, www.dei.unipd.it/∼silletti/enoch/npri.pdf, 2009

[15] P. L. Rosin, Measuring shape: ellipticity, rectangularity, and triangularity, *Machine Vision and App.*, 14(3):172-184, 2003

[16] P. Schmid, Segmentation of Digitized Dermatoscopic Images by Two-Dimensional Color Clustering Comparison, *IEEE Trans. on Medical Imag.* 18(2):164-171, 1999

[17] A. M. Turing, Computing machinery and intelligence,*Mind*, 59(433-460), 1950

[18] R. Unnikrishnan, C. Pantofaru and M. Hebert, Toward Objective Evaluation of Image Segmentation Algorithms, *IEEE Trans. Pat. Anal. Mach. Intel.*, 29(6):929-944, 2007

[19] R. Unnikrishnan and M. Hebert, Measures of Similarity, *IEEE Workshop Appl. Comp. Vis.*, 394-400, 2005

[20] H. Zhou et al. Feature-Preserving Artifact Removal from Dermoscopy Images, *Proc. SPIE* , 6914, 2008

[21] H. Zhou et al., Spatially constrained segmentation of dermoscopy images, 2008, *5th IEEE ISBI*, 800-803, 2008