

# Validation of segmentation techniques for digital dermoscopy

Guillod Joel<sup>1</sup>, Schmid-Saugeon Philippe<sup>2</sup>, Guggisberg David<sup>1</sup>, Cerottini Jean Philippe<sup>1</sup>, Braun Ralph<sup>1</sup>, Krischer Joakim<sup>1</sup>, Saurat Jean-Hilaire<sup>1</sup> and Kunt Murat<sup>2</sup>

<sup>1</sup> University Hospital Department of Dermatology of Western Switzerland, (Lausanne and Geneva), DHURDV, Switzerland and <sup>2</sup> Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland

**Purpose:** This study aims at evaluating two automatic contour detection techniques especially developed for dermoscopic images.

**Methods:** Twenty-five images of lesions with a fuzzy boundary have been randomly selected. Five dermatologists experienced in dermoscopy have manually drawn the border of all the lesions and repeated the procedure after two and four weeks. The ability of a dermatologist to reproduce its own results was evaluated by measuring the non-overlapping area enclosed by its three successive contours. The interobserver variability evaluated the contour accuracy when using automatic or manual drawings. The mean probability that a pixel has been misclassified was computed for every observer and automatic technique.

**Results:** Experts in dermoscopy are not able to reproduce measurements precisely and the two automatic techniques

had a lower missclassification probability than those obtained by each dermatologist.

**Conclusion:** This study demonstrates that a single dermatologist should not be used as a reference, and subjective validation of lesion contour is inaccurate outside an experts' group. It is argued that image processing techniques for computer-aided diagnosis must show the best compromise within such a group.

**Key words:** computer-aided diagnosis – digital dermoscopy – image processing – malignant melanoma

© Blackwell Munksgaard, 2002

Accepted for publication 16 January 2002

**D**ERMOSCOPY, ALSO named epiluminescence microscopy, or skin surface microscopy, is now widely recognized to enhance the performance of the clinical diagnosis of pigmented skin lesions (1, 2). A new semiology (3) based on this technique defines the general features of pigmented skin lesions (e.g. shape, asymmetry, colour distribution) and specific patterns like the pigmented network, pseudopods or globules. Dermoscopy is therefore, a new tool, which assists dermatologists in the early diagnosis of malignant melanoma. The diagnosis procedure is summarized in Table 1.

The first types of images that have been used for pigmented skin lesions diagnosis were clinical macroscopic images. A number of diagnosis features have been evaluated for such images and the results have led to the development of diagnosis schemes (4). However, the analysis of pigmented skin lesions through macroscopic images is

limited because there is almost no structural and colourimetric information. Dermoscopic images, where an oil immersion is used to render the epidermis translucent, have given a new dimension to skin cancer diagnosis. This microscopy reveals all the pigmented structures with their different colour shades, depending on the pigment depth, and a diffuse limit between the lesion and the healthy skin. In order to perform wide population screening and to enhance the clinical approach of image processing techniques for digitized macroscopic and dermoscopic images are now being developed by some authors (5–18).

Because the information content of dermoscopic images is much more complex than that of macroscopic images, the visual evaluation of diagnostic features has become a difficult problem for which efficient and therefore complex image processing techniques must be developed. The first

TABLE 1. Dermatologist's approach of the pigmented skin lesion

Step	Method	Result
1. Clinical observation	Patient history Inspection and palpation Clinical ABCDE rule	Clinical image as observed by naked eye, with liquid film Initial impression and risk factors
2. Dermoscopy		
a. Image acquisition	Dermoscope (10x), stereomicroscope (6–40x), digital microscope	Dermoscopic image
b. Dermoscopic features identification	Experienced clinician Computerized image processing	Global and local features
3. Evaluation of dermoscopic features	The two steps procedure (18)	
a. First step	Diagnostic algorithm to differentiate between melanocytic and non-melanocytic lesions	Non-melanocytic lesion diagnosis
b. Second step	Differentiation between benign melanocytic lesions and melanoma: ABCD rule Menzies' scoring method 7-point checklist Pattern analysis Mathematical classifiers (computer)	Probability indexes for melanoma
4. Lesion and patient management	Experience and knowledge Image comparison	Action: no further examination, clinical or digital follow-up, excision

step in any image processing system is the image acquisition (Table 1). Then the detection of the lesion border inside the image is a mandatory step for the extraction and quantification of features, either automated using processing or evaluated by human. This is a very important step because any error at this stage would of course bias all the subsequent measurements and would, therefore, reduce the accuracy of the final result. In a clinical setting, the lesion border is mentally evaluated by the dermatologist when he/she diagnoses a pigmented skin lesion under the dermoscope. This evaluation is subjective and will influence the quality and reproducibility of some diagnostic criteria, such as the symmetry, the border regularity, and the size of the lesion.

During the last years, many techniques for contour detection have been investigated by the Swiss Federal Institute of Technology in Lausanne and the University Hospital Department of Dermatology of Western Switzerland (11–15).

Preliminary results have shown the accuracy of two methods based either on image segmentation or on contour detection. The question then is, how far can one assume that such algorithms are adapted to a classification system and this problem of validation has not been yet addressed in the literature. While several studies are dealing with computer aided diagnosis systems for skin cancer, none of them focuses on the extraction of the parameters and their significance. Moreover, no

profound justification has been made on the selected extraction methods. The validation process is usually applied to the final results of the classification system but not to the internal algorithms which compose it. Because there is no objective way to define what is a valid algorithm, there is a need to find a compromise. Several contour detection methods would be candidates, as we cannot evaluate all of the existing possibilities. The two algorithms we have developed, fulfil criteria for a practical system: execution speed, no aberrant results on hundreds of images, low sensitivity to image 'quality' variations, and the use of objective and constant criteria. The current investigation should help to understand the extent and the limitations of the human evaluation in the selection of image processing algorithms. This problem is especially sensitive when dealing with dermoscopic images, where the goal is not to mimic the physician but to avoid subjective evaluation in order to get reproducible results.

## Objectives

The aim of this study is to evaluate the ability of dermatologists to draw the contour of pigmented skin lesions and to reproduce their results, and to assess the intra and/or intervariability of the solutions among other dermatologists. This intra- and interreproducibility will be compared with automated solutions calculated by different

image processing techniques. The results should be used to validate segmentation algorithms which detect pigmented skin lesions in dermoscopic images.

## Materials and Methods

### *Selection of Images*

All the lesions were taken from patients in our department. They were photographed with the Heine Dermaphot (TM) (Heine Optotechnik, Kientalstrasse 7, D-8036 Herrsching) after application of immersion oil. We have used Fujichrome Sensia II 100 ASA films that have been all processed in the same laboratory. Then, the slides were digitized with the standard Kodak PhotoCD technique. The sizes of the images used in this study were  $768 \times 512$  pixels, 16 mio colors (24 bits). Twenty-five digital images have been chosen from a set of four hundred images of pigmented skin lesions. An engineer selected the images that he obviously considered as having an ill-defined border at least on part of the lesion. This criterion was retained because it would be easier for any human observer to draw the border of sharp lesions. For practical reasons, all the images were cropped at  $512 \times 512$  pixels under the condition that each lesion was fully included in the image.

### *Human observers*

Five different observers have independently drawn the borders of the lesions. They were practising dermoscopy during at least 3 years and up to 8 years. A new layer has been added to each image using the software Photoshop (version 5 from Adobe Systems Incorporated). Original images were left unmodified and each observer had their own set of files and did not have access to images of the other observers. Images were displayed on high quality screen with 24 bits colour. To draw the border of the lesion, the 'pencil' tool was used on the added layer. It was forbidden to use the 'magnetic pencil' tool. This way, a polygon was drawn by the observer and he could move, add or delete points of the polygon until it corresponded to what best defined his understanding of the contour of the lesion. This polygon was saved into a file and used for further analysis. The operation was repeated for each of the 25 images. At least two weeks after the previous observation, the manipulation was

repeated a second and a third time on new sets of files of the same images and without access to the previous drawings. At the end of these three observations (minimum duration was four weeks), three different contour sets were obtained for every image and for every observer. The different contour sets have been labelled 'contours set 1', 'contours set 2' and 'contours set 3'.

### *Computerized contour detection*

Separation between the lesion and surrounding skin were obtained by image segmentation and by contour detection. Both approaches have been developed recently (11–13). The clustering technique and the diffusion technique were applied to the 25 images of the lesion border. Of course this automatic processing is reproducible and was therefore not repeated by contrast to the human observation.

### *Divergence calculation*

In order to evaluate the variation of the contours drawn, the following measures have been used:

- Take the contour sets pair-wise (i.e. contour sets 1 + 2; contour sets 1 + 3; contour sets 2 + 3) and the three sets together (contour sets 1 + 2 + 3).
- For each observer and pair compute two masks for each lesion, one containing all pixels that have been labelled as being in the lesion at least once, and a second one containing all the pixels that have been labelled as being in the lesion only once.
- Compute the ratio between the number of pixels in the second mask and the number of pixels in the first mask, which is then considered as the divergence or drawing error. This value is then equal to zero only when the contours are identical.

### *Pixel misclassification probability*

To assess the variations between the different observers a probability image is computed for every lesion based on the different contour results, including those obtained with our two automatic techniques. Such a representation can be used to compare one contour with the others. Only pixels that have a non-zero probability to be inside the lesion were considered. Then, for every contour the mean probability that a pixel has been misclassified is calculated and named the pixel misclassification probability. This measure allows to state if an observer has consistent results

compared to the group of observers (interobserver comparison). The probability is computed as follows:  $p(i, j) = n(i, j)/N$ , where  $N$  is the number of observations ( $N = 21$ ) and  $n(i, j)$  the number of times pixel  $(i, j)$  has been selected as being inside the lesion.

#### *Box and whisker plots*

Results are presented as box and whisker plots (19). The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. They are computed as a fraction (here 1.5) of the difference between the upper and lower quartiles. However, the whiskers limits cannot extend beyond the smallest and highest values. Outliers (+) are data with values beyond the ends of the whiskers. Finally, the boxes are notched. Notches represent a robust estimate of the uncertainty about the means for box to box comparison. They are computed from the median value at a distance equal to a fraction of the difference between the upper and lower quartiles normalized by the square root of the data size.

## Results

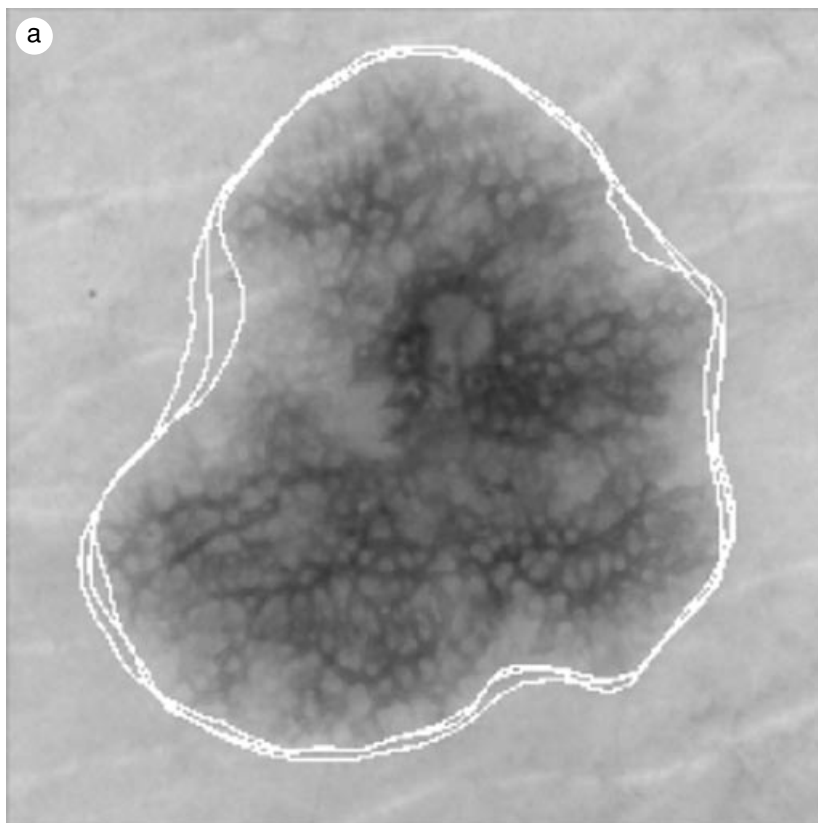
Within a period of two months, all observations were completed by the five dermatologists. From the files obtained from each dermatologist, the polygons data defining the borders were extracted. Figure 1 shows samples of manual drawing of the same lesion obtained from two dermatologists.

#### *Subjective evaluation*

The visual assessment of the contours drawn by dermatologists (Fig. 1) reveals that the contour location is uncertain in regions where the transition between lesion and healthy skin is very smooth and where the contour is non-convex (i.e. regions where the contour penetrates into the lesion).

#### *Intra-observer divergence*

For each observer, divergence ratios were calculated for each pair of drawings. Figure 2 shows the box plots corresponding, respectively, to the most (a) and the least (b) constant observers in the contour drawing. The three other observers obtained similar results (not shown).



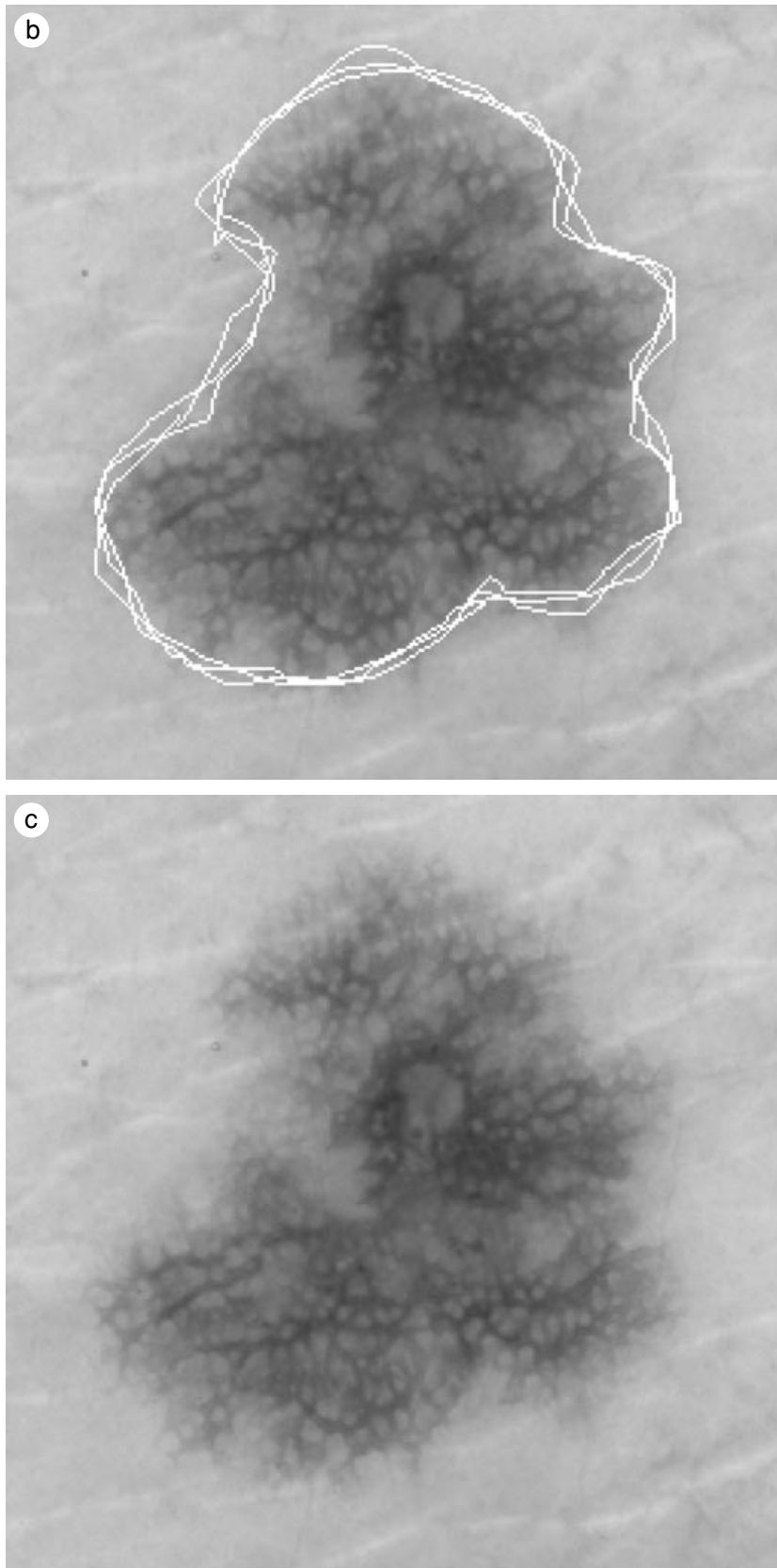


Fig. 1. (a) Contour samples drawn by physician A. The contour drawing was repeated after two and four weeks. (b) Contour samples drawn by physician B. (c) Original image.

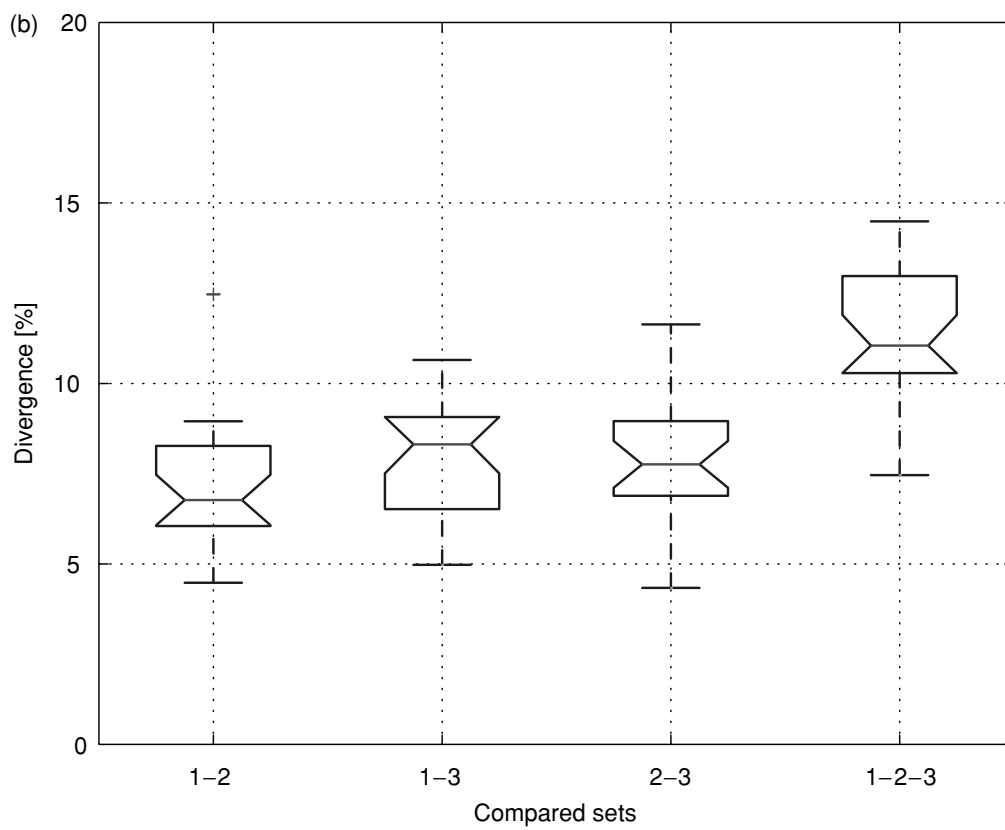
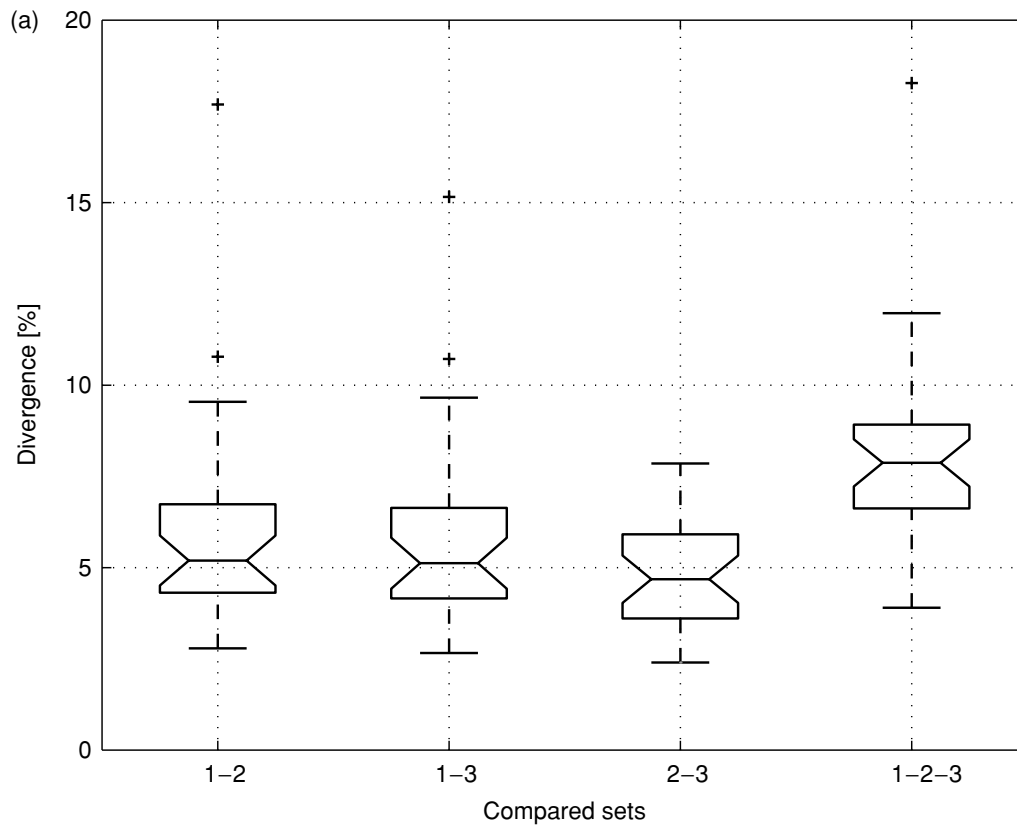


Fig. 2. (a) Box plots of the different divergence rates obtained for the three border drawings performed on the 25 images-set at different days for observer A. (b) Same for observer B.

### *Inter-observers divergence*

The probability image (example shown in Fig. 3) has been calculated with the 21 contour sets: three contour sets for each of the five observers and the two computerized contour detection techniques. In order to give the same weight to the automatic techniques than to the dermatologists, their contour results have been taken three times.

Figure 3b shows the box and whisker plots of the pixel misclassification probability obtained for our two automated techniques. The clustering technique seems to give closer results to those of the human observers than the diffusion technique.

Figure 3c and 3d shows the pixel misclassification probability of each contours set obtained from the observers A and B. In both cases the box plots show that the misclassification error is generally higher than that obtained with the automatic techniques. Interestingly, dermatologist A who was the best to reproduce his results (i.e. the lowest divergence rate, Fig. 1), has the highest misclassification probability when compared to the others.

## Discussion

In this study, we investigated the validation of methods for contour detection of pigmented skin lesion obtained by digital dermoscopy. Lesion detection is the first and mandatory processing step of a computer-aided system and can be obtained through contour detection or image segmentation techniques. Errors at this step would be carried over the succeeding measurements and could impair the final mathematical classification of pigmented skin lesions.

While visual assessment by trained dermatologists has shown that both segmentation and contour detection methods we use (12) lead to results that fit the perceived regions and contours, a more rigorous experiment has been performed in order to validate these methods. When considering the results of all three contour sets together (Fig. 2, contours set pair 1-2-3) the divergence is higher than those of pairs with two contour sets. This means that the drawing variation is different between the three successive observations and is therefore cumulated. The immediate and expected conclusion is that hand drawn contours in dermoscopic images do not show sufficient consistency to be used as absolute references (i.e. gold-standard) for the validation of image processing techniques.

Observer A was able to reproduce quite precisely his contour drawings (Fig. 1a) but obtained the highest misclassification probability (Fig. 3c) when compared to the whole group of observers. This may be explained because dermatologist A used different subjective criteria than the other observers. Finally, one cannot infer that observer A is a good reference because he is constant, or a bad one because his results diverge from that of others

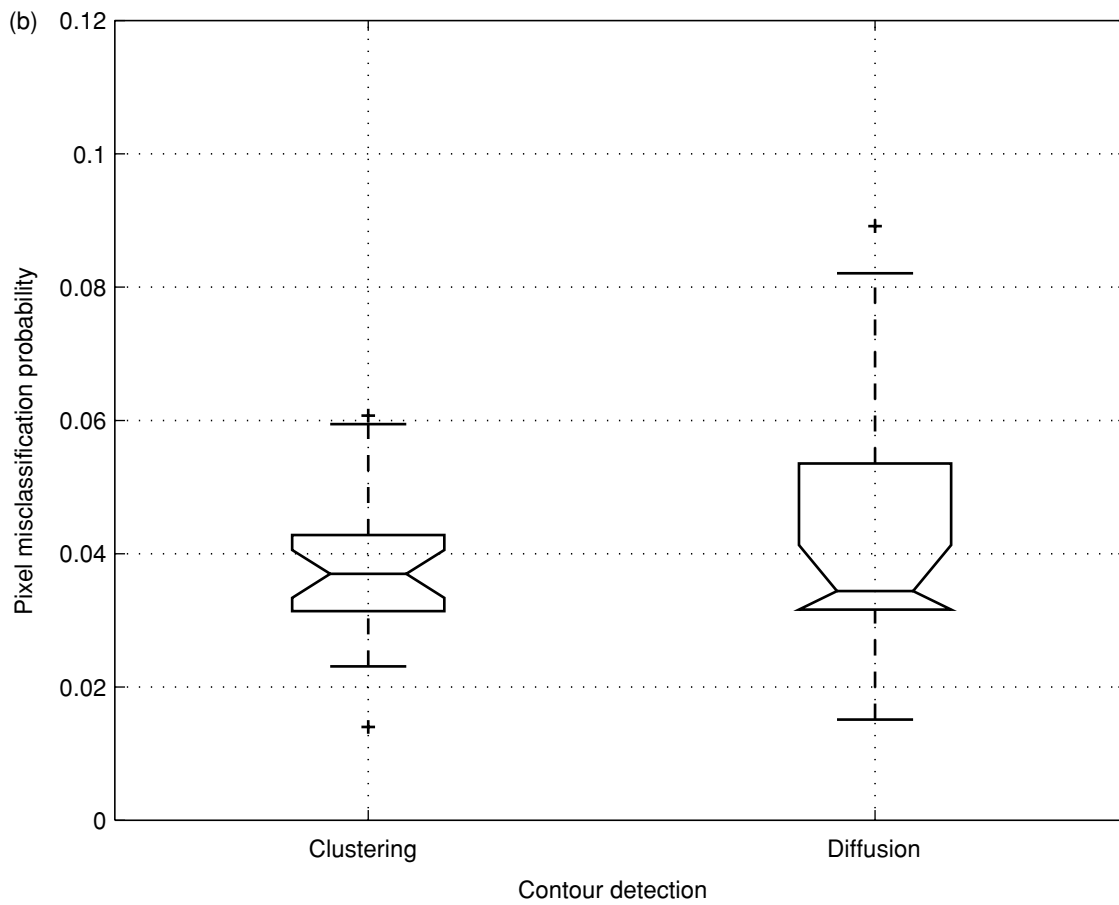
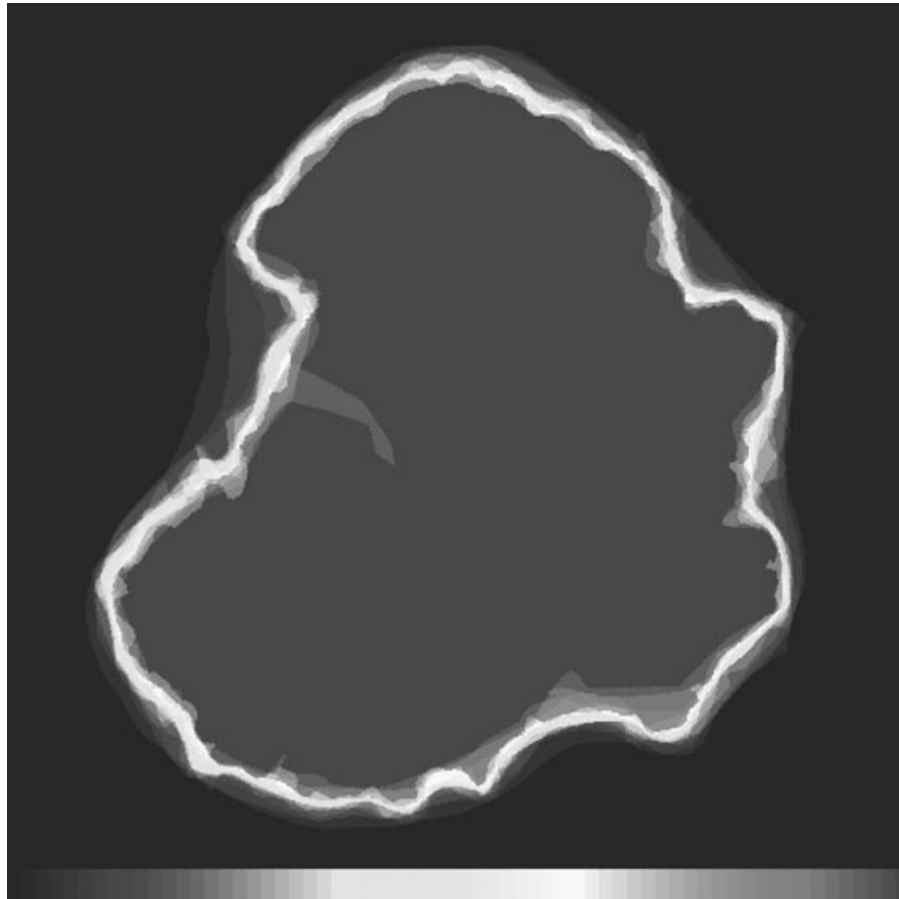
While the diffusion technique for contour detection has a low median value of misclassification error (Fig. 3b), the mean uncertainty is quite large and the error can become quite high in some cases. On the contrary, the clustering technique has almost always a low mean misclassification error (Fig. 3b), which means that it gives the best fitting results with regard to the group of experts. This result can be explained by the fact that the clustering technique uses similar criteria than the dermatologist, in that, it attempts to detect colour classes, while the diffusion technique does not draw the contour by local evaluation but instead chooses between different contour candidates based on a global measure. This can lead to larger divergency from the results provided by physicians.

In some situations where the border was very smooth, dermatologists had to draw the border line as a compromise between what appeared obviously to be the lesion and the surrounding skin. One could state that such information is lost with automated contour detection. Conversely, the delicity of the border (i.e. its sharpness or smoothness) is a distinct image measurement which can be processed in a next step, is not addressed by this study.

## Conclusions

The validation of image processing techniques for contour detection using a human reference is questionable. Actually the goal of developing a computer-aided classification system for skin cancer is to avoid human subjectivity in the processing of specific tasks. We have shown that a dermatologist, even when trained in dermoscopy, cannot be used as an absolute or gold-standard reference. Human beings are usually not able to reproduce measurements precisely and the comparison can therefore only be done with a group of experts. In that case, the behaviour of the developed computer-aided diagnosis techniques

(a)





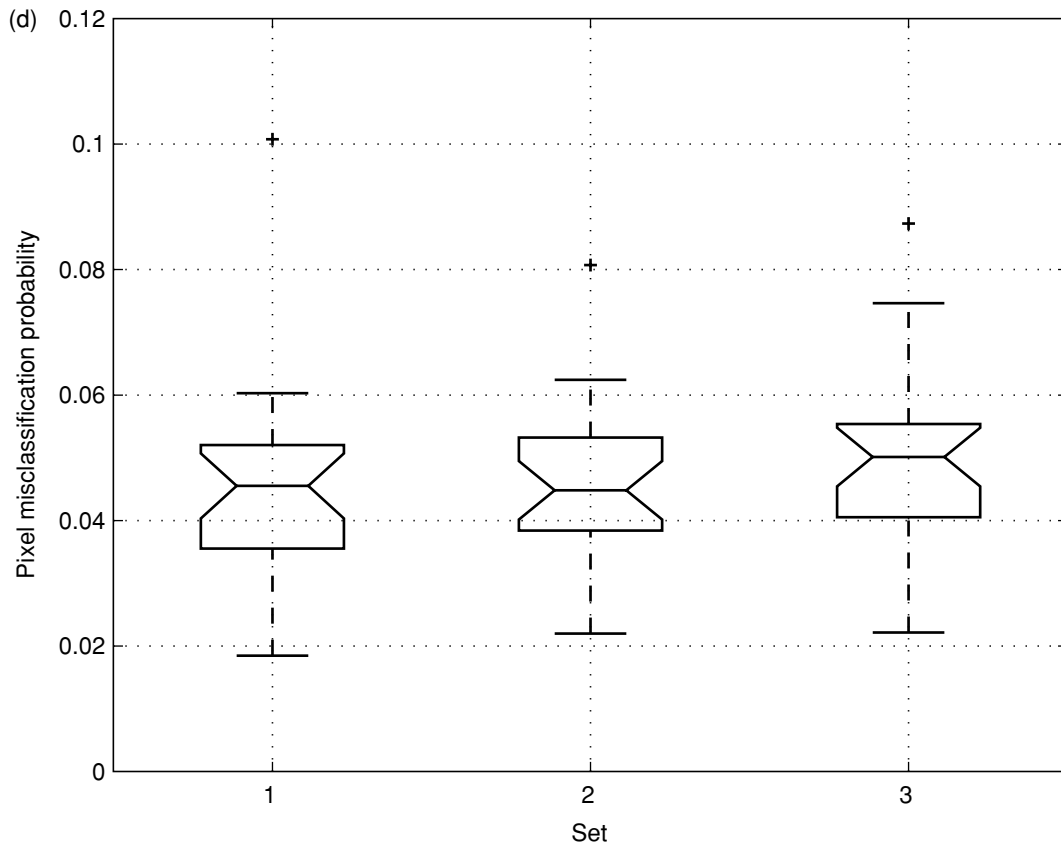
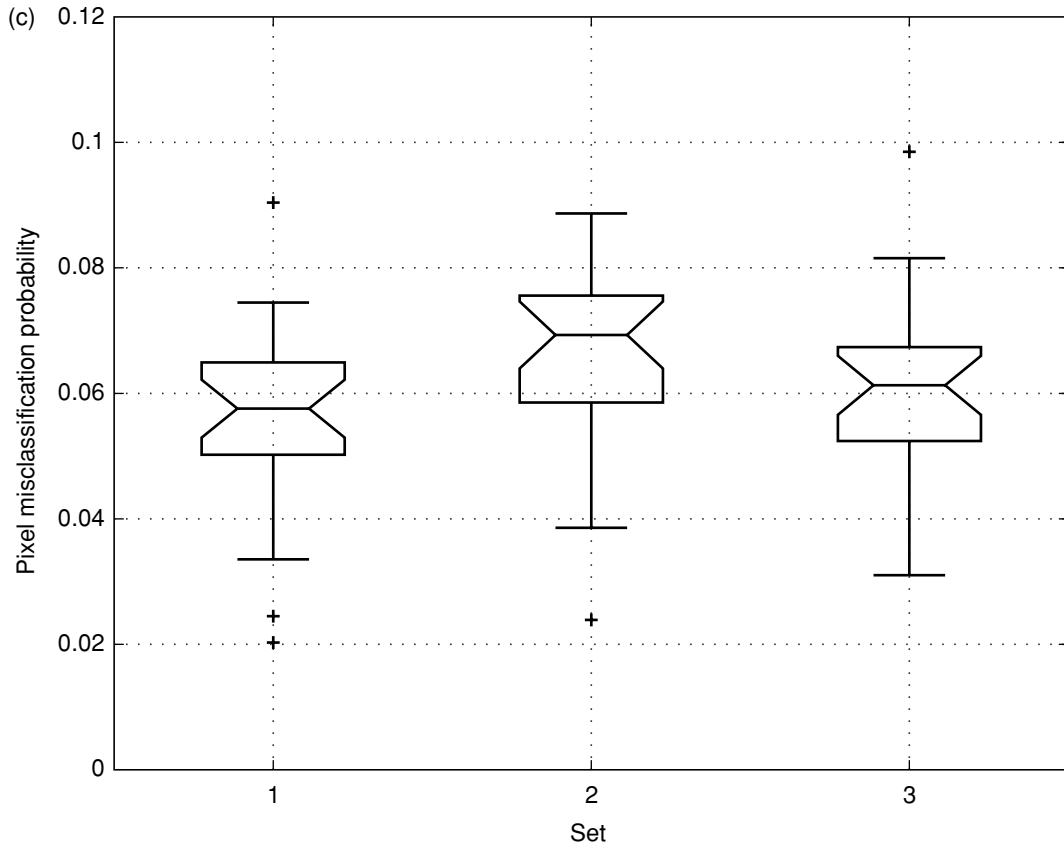


Fig. 3. Pixel misclassification probability obtained from the hand drawn and automated contours. (a) Probability image. The probability for a given pixel to be inside the lesion is 1 when shown dark red and 0 when shown blue (colour scale at the bottom from 0 to 1). (b) Probability that an image pixel has been misclassified using our computerized automated schemes. (c) Probability that an image pixel has been misclassified for observer A. (d) Probability that an image pixel has been misclassified for observer B.

must be a good compromise within the group. We could show that the within-group error introduced by the automatic schemes, and especially that of the clustering scheme, is generally lower than that introduced by the different physicians taken alone.

In this work, visual assessment has been performed as one of the ways to validate the segmentation and contour detection results, and another way will be the ability to provide the later feature extraction schemes with a lesion mask that do not corrupt the final classification. It is therefore, too early to conclude if the two techniques we have developed for contour detection are well adapted to the computer-aided classification of malignant melanoma. However, considering the above results, the provided techniques are truly serious candidates for such a system.

This study is the first publication which deals with early validation of automatic techniques for digital dermoscopy. Such a beforehand step appears to be essential in the development of a robust computer-aided classification system. Indeed, researchers need criteria in order to select reliable algorithms among the dozens which are available for each step of image processing. This study is expected to have shown a new approach in the development of innovative methods to improve computer-aided extraction and quantification of diagnostic features.

## Acknowledgements

This work has been partly supported by the Swiss National Science Research Foundation, grant #3252-53175.97. We would like to thank Dr. Laurent Lee who has corrected the english syntax of the manuscript.

## References

1. Pehamberger H, Binder M, Steiner A, Wolff K. In Vivo Epiluminescence Microscopy: Improvement of Early Diagnosis of Melanoma. *J Invest Dermatol* 1993; 100: 356S–362S.
2. Stolz W, Braun-Falco O, Landthaler M, Bilek P, Cognetta AB. *Color Atlas of Dermoscopy*. Blackwell Science, 1994.
3. Bahmer FA, Fritsch P, Kreis J et al. Diagnostische Kriterien in der Auflichtmikroskopie. Konsensus-Treffen der Arbeitsgruppe Analytische Morphologie der Arbeitsgemeinschaft Dermatologische Forschung, 17 November 1989 in Hamburg. *Hautarzt* 1990; 41: 513–514.
4. Sober AJ, Fitzpatrick TB, Mihm MC Jr. Primary melanoma of the skin: recognition and management. *J Am Acad Dermatol* 1980; 2: 179–197.
5. Stolz W, Harms H, Aus HM, Abmayr W, Braun-Falco O. Macroscopic diagnosis of melanocytic lesions using colour and texture image analysis. *Abstracts* 1990; 95: 491.
6. Green A, Martin N, McKenzie G, Pfitzner J, Quintarelli F, Thomas BW, O'Rourke M, Knight N. Computer image analysis of pigmented skin lesions. *Melanoma Res* 1991; 1: 231–236.
7. Claridge E, Hall PN, Keefe M, Allen JP. Shape analysis for classification of malignant melanoma. *J Biomed Eng* 1992; 14: 229–234.
8. Schindewolf T, Stolz W, Albert R, Abmayr W, Harms H. Comparison of classification rates for conventional and dermoscopic images of malignant melanoma and benign melanocytic lesions using computerized colour image analysis. *Eur J Dermatol* 1993; 3: 299–303.
9. Schiffner R, Stolz W, Pillet L, Harms H, Schindewolf T, Landthaler M, Abmayr W. ADAMS: a PC based acquisition, data management, and surveillance system for pigmented skin lesions using dermoscopy (Abstract). *Eur J Dermatol* 1994; 103: 414.
10. Gutkowitz-Krusin D, Elbaum M, Szwaykowski P, Kopf AW. Can early malignant melanoma be differentiated from atypical melanocytic nevus by in vivo techniques? Part II. Automatic machine vision classification. *Skin Res Technol* 1997; 3: 15–22.
11. Schmid P. Lesion detection in dermoscopic images using anisotropic diffusion and morphological coding. In *Proceedings of the International Conference on Image Processing (ICIP)*, Kobe, Japan, 24–28 October 1999.
12. Schmid P. Segmentation and symmetry measure for image analysis: application to digital dermoscopy. PhD Thesis no. 2045, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, 1999.
13. Schmid P. Segmentation of dermoscopic images by 2D color clustering. *IEEE Transactions on Medical Imaging*, February 1999; 18: 164–171.
14. Schmid P. Symmetry axis computation for almost-symmetrical and asymmetrical objects: application to pigmented skin lesions. *Med Image Processing* 2000; 4: 269–282.
15. Guilloid J, Schmid P, Agache P et al., eds. *Physiologie de la Peau et Explorations Fonctionnelles Cutanées*. Cachan Cedx: Editions Médicales Internationales, 2000; 59–73.
16. Lee T, Stella Atkins M. New approach to measure border irregularity for melanocytic lesions, in *SPIE Medical Imaging*, San Diego, 2000.
17. Donadey T, Serruys C, Giron A, Aitken G, Vignali J, Triller R, Fertl B. Boundary Detection of Black Skin Tumors Using an Adaptive Radial-based Approach, in *SPIE Medical Imaging* 2000, San Diego, 2000.
18. Soyer HP, Argenziano G, Chimenti S, Menzies SW, Pehamberger H, Rabinovitz HS, Stolz W, Kopf AW. *Dermoscopy of pigmented skin lesions: an atlas based on the consensus net meeting on dermoscopy*, 2000. EDRA Medical Publishing & New Media, February 2001, ISBN 88-86457-42-1.
19. Chase W, Brown F. *General Statistics*. John Wiley & Sons, 1992.

Address:

Joel Guilloid  
Faubourg du Lac 7  
2000 Neuchâtel  
Switzerland

Tel: + 41 32 721 29 29

Fax: + 41 32 721 29 30

e-mail: joel.guilloid@net2000.ch