Automated Delineation of Dermal-Epidermal Junction In Reflectance Confocal Microscopy Image Stacks Of Human Skin

Sila Kurugol, Kivanc Kose, Brian Park, Jennifer G Dy, Dana H Brooks, Milind Rajadhyaksha

Cite this article as: Sila Kurugol, Kivanc Kose, Brian Park, Jennifer G Dy, Dana H Brooks, Milind Rajadhyaksha, Automated Delineation of Dermal-Epidermal Junction In Reflectance Confocal Microscopy Image Stacks Of Human Skin, *Journal of Investigative Dermatology* accepted article preview 3 September 2014; doi: 10.1038/jid.2014.379.

Received 14 April 2014; revised 25 July 2014; accepted 7 August 2014; Accepted article preview online 3 September 2014

Automated Delineation of Dermal-Epidermal Junction In Reflectance Confocal Microscopy Image Stacks Of Human Skin

Sila Kurugol[a],†, Kivanc Kose[b],†*, Brian Park[c], Jennifer G. Dy[d],‡, Dana H. Brooks[d],‡, Milind Rajadhyaksha[b],‡

[a] Department of Radiology, Boston Children's Hospital and Harvard Medical School, Boston, MA

[b]Dermatology Service, Memorial Sloan Kettering Cancer Center, New York, NY

[c]NYU School of Medicine and NYU Department of Radiology, New York, NY

[d]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA

†equal contribution (shared first authorship)

‡equal contribution (shared senior authorship)

*corresponding author: Dermatology Service, Memorial Sloan Kettering Cancer Center, +1 212 6100831, kosek@mskcc.org

Abstract

Reflectance confocal microscopy (RCM) images skin non-invasively, with optical sectioning and nuclear-level resolution comparable to that of pathology. Based on assessment of the dermal-epidermal junction (DEJ) and morphologic features in its vicinity, skin cancer can be diagnosed in vivo with high sensitivity and specificity. However, the current visual, qualitative approach for reading images leads to subjective variability in diagnosis. We hypothesize that machine learning-based algorithms may enable a more quantitative, objective approach. Testing and validation was performed with two algorithms that can automatically delineate the DEJ in RCM stacks of normal human skin. The test set was composed of 15 fair and 15 dark skin stacks (30 subjects) with expert labellings. In dark skin, in which the contrast is high due to melanin, the algorithm produced an average error of 7.9±6.4μ*m*. In fair skin, the algorithm delineated the DEJ as a transition zone, with average error of 8.3±5.8μ*m* for the epidermis-to-transition zone boundary and 7.6±5.6μ*m* for the transition zone-to-dermis. Our results suggest that automated algorithms may quantitatively guide the delineation of the DEJ, to assist in objective reading of RCM images. Further development of such algorithms may guide assessment of abnormal morphological features at the DEJ.

Key words: confocal microscopy, human skin, dermal-epidermal junction, image analysis, texture analysis, machine learning.

1 Introduction

Reflectance confocal microscopy (RCM) is a non-invasive imaging technique used to examine skin. Its optical sectioning (1−3μ*m*) and nuclear-level resolution (0.5−1.0μ*m*) are comparable to that of pathology. Stacks of en-face images are routinely acquired to examine skin in depth (100−200μ*m*), and mosaics near the dermal-epidermal junction (DEJ) to examine in lateral extent (10*mm*×10*mm*). Basal cell carcinomas have been diagnosed in vivo with 92−100% sensitivity and 97−85% specificity and melanomas with 92−88% sensitivity and 70−84% specificity (Guitera et al., 2012; Nori et al., 2004). Initial implementation in academic clinical settings (Alarcon et al., 2013; Pellacani et al., 2014) showed that RCM imaging combined with dermoscopy, reduced the number of unnecessary biopsies and thus also the economic burden associated with skin cancer management. This success has been achieved by a small cohort of "early adopter" clinicians, who, through working with the technology and performing the clinical studies, have become experts in reading RCM images. However, RCM images are more challenging to read than pathology. The imaging is in *en face* orientation (instead of orthogonal), and with only one source of contrast (reflectance), instead of two, as with hematoxylin and eosin. Consequently, the images appear in gray-scale contrast (instead of purple and pink colored). In addition, the contrast and signal-to-noise varies with pigmentation conditions and degrades with depth, especially below the DEJ.

Thus, the ability to analyze RCM images for diagnosis is currently confined to the early adopter clinicians. For the larger cohort of novice (interested in, but new to RCM) clinicians, the difficulty of reading images is a barrier against training and wider adoption of this technology. To address the need for training, machine learning-based image analysis is being investigated to provide quantitative and objective approaches for reading images. (Gareau et al., 2010; Koller et al., 2011; Kurugol et al., 2011; Wiltgen et al., 2008) One of the first studies reported a method based on texture analysis for automated identification of diagnostically significant regions in RCM images of melanocytic lesions(Koller et al., 2011; Wiltgen et al., 2008). Another group of researchers developed a method to automatically quantify the spread of pagetoid melanocytes in epidermis and disarray at the DEJ level in order to detect superficial spreading melanomas (Gareau et al., 2010). Subsequently, our group carried out feasibility studies on a texture analysis approach to automatically delineate the DEJ, in terms of an epidermis-to-dermis transition zone, in RCM stacks of fair skin in vivo (Kurugol et al., 2011). The boundaries of the transition zone were localized with an average error, on a small set of four image stacks, of ∼8.5±6.8μ*m*, with epidermis versus dermis classification rates above 85%.

The rationale for addressing localization of DEJ is that this junction and its vicinity, in which the majority of diagnostically important features are found, is routinely examined in pathology. This was accomplished in most of the clinical studies by first localizing the DEJ in RCM stacks and then acquiring (and analyzing) RCM mosaics at, just above, and just below the junction (Alarcon et al., 2013; Gill et al., 2013; Pellacani et al., 2007). This approach is now standard practice for imaging on patients. When visually examining RCM stacks, expert readers typically use texture and contrast differences between layers of the epidermis and dermis, in order to locate the DEJ. For example, granular layers characteristically appear as honeycomb patterns formed by polygonal cells, with dark nuclei surrounded by bright grainy cytoplasm, whereas spinous and basal layers appear in a distinct cobblestone pattern. The papillary dermis appears different, as a dark band between the epidermis and the underlying relatively brighter reticular dermis, sometimes including dark lumen-like structures corresponding to capillary loops (occasionally containing bright blood cells). Furthermore, in dark skin, due to high concentration of melanin, the basal layer appears with distinctly bright contrast, which makes it possible to locate the DEJ more reliably than in fair skin. However, the current visual approach for delineation of the DEJ in RCM stacks is subjective and produces significant variability. We hypothesized that machine learning-based algorithms may provide a more quantitative, objective approach and performed an initial investigation in image analysis methods (Kurugol et al., 2011).

In this article, we expand upon our previously reported approach, to include an algorithm for DEJ localization in dark skin (Section 4), and report the results of testing on an extended dataset, with validation against "ground truth" segmentation of epidermis versus dermis by expert readers (Section 2).

2 Results

In Table 1, a summary of the outcomes, showing the best, worst, and average results for DEJ delineation in dark and fair skin is presented. The complete table of results for all of the 15 stacks of dark and 15 of fair skin is included in the supplementary material.

We report the mean and the standard deviation of the error, as well as the percentage of tiles that are within 15μ*m* error, averaged across all 15 stacks for each type of skin, as well as for the best and the worst cases. The table also reports the classification accuracy, in the form of "confusion matrices", for the two layers (epidermis, dermis) in the

case of dark skin and three (epidermis, transition zone, dermis) in the case of fair skin. Specifically, in the confusion matrices, diagonal elements show percent correct classification, while the off-diagonal elements show the percentages that were misclassified. The table shows that, in dark skin, the epidermis was correctly classified in 89% of the tiles and the dermis in 87%. The epidermis-as-dermis misclassification was 11% and dermis-as-epidermis was 13%. In fair skin, the confusion matrix indicates that algorithm was correctly classified in 64%, 41%, and 75% of tiles for epidermis, transition region, and dermis, respectively. The dermis and epidermis were well distinguished from each other on the other hand, with small percentages of epidermis-dermis misclassification.

To illustrate our results (Figs. 2-4), we show three-dimensional visualizations of example results from selected stacks comparing algorithmic and expert delineations. Orthogonal visualizations of *en face*, sagittal, and coronal slices are shown from example stacks of both dark (Fig.2) and fair (Fig. 3) skin. In both figures, the cross-hairs on the *en-face* views show the locations of the corresponding sagittal (red in dark, orange in fair skin stacks) and coronal (green in dark, and light blue in fair skin stacks) slices. Expert boundaries are shown with light blue lines and algorithm boundaries in orange on all slices. The sagittal and coronal views are analogous to the standard orientation of histology sections and illustrate the anisotropic resolution of the RCM imaging. Fig. 4 shows three-dimensional surface views of the boundaries in example stacks for both dark and fair skin. Fig. 4-a shows the algorithmic delineation of the DEJ in dark skin, with color mapped onto the surface to show the separation error between this boundary and the labeling of the experts. In Fig. 4-b, the surfaces show the algorithmic segmentation of the epidermis-to-transition boundary (in color in the upper portion) and the dermis-to-transition boundary (color in the lower portion) in fair skin. Again, error between the expert and algorithm is color mapped on the surfaces.

3 Discussion

Over 30 stacks, we achieved an average error of ≤8.5μ*m*. Examination of results across all 15 dark skin stacks revealed that for 9 of them, the error was less than 15μ*m* for 86% or better of the tiles. A small subset of those stacks had markedly worse performance; nonetheless the mean error was less than 15μ*m* for all 15 stacks and only 5 stacks had mean error >8μ*m*. The stacks with less accurate performance consistently had both higher mean error and higher standard deviations, suggesting that a subset of tiles in those stacks were particularly problematic. We found that, in those cases, brightness contrast from highly melanized basal cells was lacking in some locations. In fair skin, 11 of both boundaries met the <15μ*m* accuracy criterion on 86% or more tiles. Although the worst case stack had mean boundary errors of ≈20 and ≈17μ*m*, only 4 upper boundaries and 4 lower boundaries had mean errors greater than 8μ*m*. When the method performed less accurately, it did so for both boundaries, and mean errors were higher than for dark skin, suggesting a more general breakdown of the method. Confusion matrix results confirmed that, in dark skin, we achieved accurate tile classification for both epidermis and dermis in almost 90% of the tiles across all stacks, with errors evenly distributed between dermis and epidermis. In fair skin, correct classification was lower, with dermis classified more accurately than epidermis and transition classified much less accurately, as confirmed by the off-diagonal matrix entries in Table 1.

With average error comparable to the thickness of the basal cell layer, we believe that we achieved our goal of DEJ identification in dark skin. In fair skin, although average results were comparable to those is dark skin, in some stacks the accuracy was low in several tiles. We believe this may reflect a combination of intrinsic difficulty and insufficiently accurate labeling in both training and ground truth. Visual similarity in fair skin between lower epidermis and papillary dermis across rete ridges, and lack of texture around and at the basal layer, in part led to less successful delineation (e.g. mismatch of expert and algorithmic borders as seen in the coronal view in Fig. 3). In addition to the intrinsic challenges posed by that visual similarity, as noted in the next paragraph limitations of expert labeling accuracy and repeatability may themselves limit performance, since parameters (e.g. for the Locally Smooth Support Vector Machine (LS-SVM) classifier) were learned from stacks whose labeling was itself uncertain (see saggital and coronal views of Fig. 4, which show wide expert-labeled transition regions in places).

Table 1 shows that epidermis/transition distinction is particularly difficult. Whereas the distinction was easy near dermal papillae, it was relatively more difficult near rete ridges due to loss of resolution and the resulting blurred appearance in the epidermal cellular patterns and the dermal collagen patterns. However, as our labelers gained experience, they reported better appreciation of subtle yet observable variations in the blurred texture between lower epidermis in the rete ridges and papillary dermis. When observed closely, the lower spinous cell layers appeared homogeneously blurred, while the underlying collagen appeared heterogeneously blurred with perceptible fibrous patterns. Therefore improved labeling may be possible, leading to improved supervised learning models, and thus to improved performance, through thinner transition boundary segmentations or, perhaps, by only identifying one (DEJ) boundary with no transition region.

Further improvement in the machine learning algorithm may be realized by using multiple stacks to train the LS-SVM fair skin model in a multi-level approach. Similarly textured RCM stacks would be grouped together to develop specialized templates and LS-SVMs. In classification, stacks could be compared to templates to determine their texture type, then classified with the appropriate LS-SVM model. Dark skin tiles with poor melanin contrast could also be identified automatically and fed into a multi-classifier algorithm. Finally, we have developed a prototype RCM "dark versus fair" skin classifier (not reported here); if verified on a larger dataset, it could allow unification of our two separate algorithms.

Summarizing, our approach may suffice for many applications. For example, clinicians currently examine RCM image mosaics covering larger areas for diagnosis. Typical acquisition is of several mosaics positioned with respect to (*i.e.* at, above, and below) a putative average DEJ depth, in turn requiring identification of that "average DEJ depth" from one or more stacks in the mosaic region. This is currently done by visual assessment, and thus is subjective, with high inter-clinician variability. Our algorithms may allow standardized imaging for both research and clinical practice. Moreover, as our approach provides the 3D structure of the DEJ, it can also be utilized to quantify its microanatomy.

Other limitations include consensus validation by only two experts. A larger study with several readers, accounting for inter-rater variability, would more precisely define our precision. Dividing stacks into fixed size tiles is somewhat arbitrary. Dependence on these divisions should be studied. Since skin morphology varies significantly across age and site, extensive further testing with larger datasets is essential. All skin imaged here was normal from healthy volunteers. Lesion abnormalities, in particular with disrupted DEJs (*e.g.* some types of malignant melanoma), are likely to provide additional challenges. We believe they can be addressed within our framework, but this topic remains to be studied.

4 Materials & Methods

Our algorithmic approach mimics key aspects of the visual approach used by expert readers. Due to the *en face* orientation of RCM images, dermal papillae and the DEJ are associated with the appearance of rings of basal cells (Fig. 1). Therefore, rather than looking for a complete DEJ boundary across an entire RCM stack, readers search for ring-like patterns of basal cells. These patterns occupy small areas within each image. Thus, the size of these areas determines a spatial resolution that is employed by the readers' visual perception during their search for the DEJ in each image. (Note that this visual resolution is entirely different and on a much larger scale than the μm-level optical resolution of the actual imaging, which is determined by the lens of the confocal microscope.) Similarly, the processing in our algorithms relies on spatial resolution implemented in the form of small square-shaped areas or "tiles" within each image. Each stack of RCM images is first divided into such tiles and then processed as a collection of non-overlapping "stacks of tiles" or "tile stacks".

Another key aspect that we mimic is the use of contrast versus texture to locate the DEJ. In particular, since the significance and detectability of these features differ between dark and fair skin, we developed separate algorithms for each. In the case of dark skin (Types III to VI), the basal layer appears bright in RCM images due to the presence and high reflectivity of melanin (illustrated in Fig. 1(a)). Thus, the basal layer can be usually localized by scrolling up and down in a stack and looking for obvious changes in contrast (intensity brightness) due to rings of basal cells. The DEJ is then delineated as the inner boundary of the ring, being an *en face* image (this would correspond to the lower boundary of the basal layer in a conventional orthogonally oriented section of pathology). Since the rings have distinguishable contrast, the spatial resolution for image processing and analysis can be as small as the size of a basal cell. Therefore, the dark skin algorithm uses $16\mu m \times 16\mu m$ tiles, approximately the size of 1−2 basal cells. However, sometimes there can be several bright regions in multiple images within a stack, due to, for example, dermal collagen. In such cases, we also use other structural or textural features to distinguish the basal layer. These textural features are discussed in detail in section 4.1 and the supplementary technical document.

In fair skin (Types I to II), the ring-like pattern of basal cells that enclose the papillary dermis appear with weak contrast. Therefore, readers cannot reliably detect the location of the basal layer. Instead, they typically utilize local textural features to delineate the DEJ. Patterns of texture in this case are larger than the size of a basal cell, such that the relationship between any observed location and the neighboring area becomes more important. Thus, our algorithm for fair skin operates on larger tiles, $25\mu m \times 25\mu m$, in order to more effectively incorporate information from texture characterization of neighboring areas. Due to the lack of contrast at the basal layer, our expert readers often cannot determine the exact location of the DEJ but rather tend to delineate a transition region between epidermis and dermis, (which would include the DEJ). Therefore, we designed the fair skin algorithm to also delineate a transition region, with two boundaries, an epidermis-transition boundary and a dermis-transition boundary.

In the rest of this section, we first describe the data acquisition and the preprocessing stages that are common for both algorithms. Then we briefly describe the technical details of each algorithm, followed by the description of a common post-processing step. Finally, we describe the error metrics used in our study.

4.1 Data Acquisition and Preprocessing

Our RCM dataset consists of 15 stacks of fair and 15 of dark skin, acquired on the forearms of 30 subjects. The imaging was performed on volunteer subjects with their written consent under an IRB-approved protocol. Stacks were determined at the time of acquisition as being from either fair or dark skin by direct observation of the subjects' skin type. The acquisition was performed with a commercial confocal microscope (Vivascope 1500, Caliber Imaging & Diagnostics, Rochester, New York), which has been routinely used in all reported clinical studies, e.g. (Alarcon et al., 2013; Gill et al., 2013; Guitera et al., 2012; Nori et al., 2004). Each image has a field of view (FOV) of $0.5mm \times 0.5mm$, with lateral resolution of $\sim 0.7\mu m$ and optical sectioning thickness of $\sim 3\mu m$. The depth-spacing between images in each stack is $\sim 1\mu m$.

In each stack, the epidermis and dermis regions were manually labeled by consensus between at least two expert readers using an open source segmentation tool called Seg3D (CIBC, 2013). The expert labeling was used as the ground truth for testing accuracy. The fair skin algorithm obtained some of its parameters by training on expert labeling of an additional stack, the one used in our earlier study (Kurugol et al., 2011). This stack was used exclusively for training purposes here in this study, and was not used for any testing.

Our automated processing starts by first registering the images in each stack in the lateral direction, , in order to correct for misalignment due to patient motion during imaging. This step is important because our algorithms rely on change in local tile-specific features with depth. Each stack was then processed tile by tile to calculate textural features. These textural features are mathematical representations of structure in epidermal and dermal layers, calculated for each pixel in the images using a set of its adjacent pixel values. We employed a large set of well-known textural features: graininess, co-occurrence matrix features, (Haralic, 1979) statistical moments,(Randen and Husoy, 1999) wavelet packet decomposition coefficients,(Laine and Fan, 1993; Randen and Husoy, 1999) log-Gabor filter features,(Field, 1987) and radial spectrum features.(Gonzales and Wood, 2002)

This large feature set contained considerable redundancy, which we minimized through a fast filter method (Yu and Liu, 2004) based feature selection process to determine the least redundant and most discriminative subset of features. The method used the Fisher class separation distance measure to find the least mutually redundant subset of features. Our fast filter based analysis suggested that log-Gabor and wavelet features were the most discriminative. These features are sensitive to spatial frequency and highlight textural differences among the layers (e.g. blurry appearance of the collagen patterns in dermis compared to the relatively sharper appearance of cellular patterns in the epidermis). The mathematical and technical details of the feature extraction procedure, properties of individual features, and the feature selection process are described in our preliminary study (Kurugol et al., 2011) and the accompanied supplementary technical document.

4.2 DEJ Delineation in RCM Stacks

Dark Skin Algorithm:

As stated, the basic motif of our dark skin algorithm is to detect intensity changes in each stack of tiles along the axial (depth) direction. Thus, we first computed the median intensity of each tile and constructed a median intensity profile as a function of depth. We smoothed the profile using a Gaussian filter (width parameter $\sim 5\mu m$ (5 images)).

In most, but not all, tile stacks, we observed single peaks in the median intensity profile, which unequivocally corresponded to the locations of the basal cell layer. At these peaks, the textural features were calculated to construct a texture template for basal cells. For the remaining tile stacks, with multiple peaks, the peak of interest was found by comparing each against this texture template and selecting the one that was most similar. Finally, for both single and multiple peak profiles, the DEJ was located at the first inflection point below the selected peaks.

Fair Skin Algorithm:

In fair skin, as mentioned earlier, intensity contrast by itself is not enough for reliable delineation of the DEJ, or even of a transition zone. To localize the transition zone boundaries, epidermis-to-transition-zone and transition zone-to-dermis, we developed a two step algorithm(Kurugol et al., 2011). We briefly summarize the algorithm here for completeness; details can be found in our previous report (Kurugol et al., 2011) and accompanied supplementary technical document.

Due to the loss of resolution with depth and the speckle noise in RCM images (resulting from the complex structure of skin), classification based purely on textural features from individual tiles was not robust. For example, the texture of lower epidermis in rete ridges and that of the papillary dermis below is particularly difficult to distinguish. By comparison, textural differences between epidermis above dermal papillae and the underlying dermis is easy to distinguish. However, the textures in any sequence of neighboring tiles in the depth direction are highly correlated within a skin layer (i.e.,when the sequence is entirely in either the epidermis or dermis). Thus, in our first step, consecutive tiles in each tile stack with similar texture appearance were grouped into intervals along the axial (depth) direction, and then fed into two binary classifiers: epidermis versus non-epidermis and dermis versus non-dermis.

To implement this grouping, we developed a sequential segmentation (SS) algorithm, which divided each tile stack into shorter sub-stacks of consecutive tiles represented with an affine model of features. The number of such sub-stacks as well as their boundary locations were determined using dynamic programming (Cormen et al., 2001). The resulting boundaries between sub-stacks of tiles were then used in the second stage of the fair skin algorithm.

In the second stage, we used the sub-stack borders as the candidates for epidermis-transition and transition-dermis decision boundaries. We first classified each tile as epidermis (dermis) vs. non-epidermis (non-dermis) using a support vector machines (SVM)(Cortes and Vapnik, 1995) based supervised learning method. In supervised learning, a machine learning model is trained on samples from a labeled set ("training set"). Then the data to be classified is fed into the trained machine learning model. We used SVM because it is one of most effective and widely used machine learning methods and, furthermore, allowed us to take advantage of a variant, called locally smooth SVM (LS-SVM).(Vural et al., 2009) LS-SVM takes into account the expected resemblance between neighboring tiles within images to increase robustness. Here, this corresponds to the expert readers' visual process of looking at local context by examining the area around each location of interest when determining the transition boundaries.

To localize the two boundaries of the transition zone, we trained two LS-SVM models. The first one classified epidermis vs. non-epidermis and operated in a top-to-bottom (epidermis-to-dermis) direction. The second classified dermis vs. non-dermis and operated bottom-to-top (dermis-to-epidermis). We fed textural features from each tile into the trained LS-SVM models to obtain probability values of belonging to either epidermis or dermis. The mean probabilities for each sub-stack of tiles (which were determined from the SS algorithm in the first stage) were calculated, based on the boundaries. Groups of tiles whose probability of belonging to the epidermis class was lower than a threshold (here, set to 0.4) were classified as non-epidermis. Non-dermis sub-stacks were determined in a similar way. The intersection of non-epidermis and non-dermis regions was taken as the transition zone. The DEJ was assumed to be within this zone, between the upper and lower transition boundaries.

4.3 Postprocessing: Final Boundary Localization

The result of the processing was the delineation of either the DEJ in dark skin or the transition zone in fair skin with a prescribed visual spatial resolution defined by the size of tiles used in the processing. For visualization purposes, we applied a Gaussian smoothing filter (support set 5×5, standard deviation 0.75, in units of tiles) to the discontinuous estimated boundary, followed by cubic spline interpolation to interpolate the boundaries from tiles to individual pixels.

4.4 Performance Metrics and Error Analysis

Three metrics were used to assess the performance of the algorithm. The first was the error, in terms of the separation between the expert-labeled and algorithm-delineated boundaries, summarized as the mean and standard deviation of the error distribution for all stacks. The second was percentage of tiles for which the algorithm-delineated DEJ was within 15μm from the expert-labeled boundary. We chose 15μm as the threshold because it is approximately the thickness of a basal cell. The third was classification/mis-classification accuracy, calculated for the epidermal and dermal layers for both dark and fair skin. To visually demonstrate the agreement between the algorithmic- and expert-segmented DEJ, we've also included several images of slices from both expert and algorithmic segmentations.

5 Conflict of interests

Dr. Milind Rajadhyaksha owns equity in Caliber ID (formerly, Lucid Inc.). Pellacani for extensive discussions and guidance. Sila Kurugol, Kivanc Kose, Brian Park, Jennifer Dy, and Dana H. Brooks do not have any conflict of interest.

6 Acknowledgments

We thank Miguel Cordova for his help in labeling the RCM stacks that we used in this study, and Dr. Giovanni Pellacani for extensive discussions and guidance.
This project was supported by grants (R01EB012466), (R01CA156773),(P41GM103545) from the NIH.

References

Alarcon, I., Carrera, C., Palou, J., and et al. (2013). Impact of in vivo reflectance confocal microscopy on the number needed to treat melanoma in doubtful lesions. British Journal of Dermatology, online:available. ISSN 1365-2133. 10.1111/bjd.12678.

CIBC (2013). Seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI), Download from: http://www.seg3d.org.

Cormen, T. H., Stein, C., Rivest, R. L., and et al. (2001). Introduction to Algorithms. McGraw-Hill Higher Education, 2nd edition. ISBN 0070131511.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3):273–297. ISSN 0885-6125. 10.1007/BF00994018.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. Journal of Optical Society of America, 4(12):2379–2394.

Gareau, D., Hennessy, R., Wan, E., and et al. (2010). Automated detection of malignant features in confocal microscopy on superficial spreading melanoma versus nevi. Journal of Biomedical Optics, 15(6):061713–061713–10.

Gill, M., Longo, C., Farnetani, F., and et al. (2013). Non-invasive in vivo dermatopathology: identification of reflectance confocal microscopic correlates to specific histological features seen in melanocytic neoplasms. Journal of the European Academy of Dermatology and Venereology, online:available. ISSN 1468-3083. 10.1111/jdv.12285.

Gonzales, R. C. and Wood, R. E. (2002). Digital Image Processing. Prentice Hall.

Guitera, P., Menzies, S. W., Longo, C., and et al. (2012). In vivo confocal microscopy for diagnosis of melanoma and basal cell carcinoma using a two-step method: Analysis of 710 consecutive clinically equivocal cases. Journal of Investigative Dermatology, 132;10:2386–2394.

Haralic, R. M. (1979). Statistical and structural approaches to texture. Proceedings of the IEEE, 67:786–804.

Koller, S., Wiltgen, M., Ahlgrimm-Siess, V., and et al. (2011). In vivo reflectance confocal microscopy: automated diagnostic image analysis of melanocytic skin tumours. Journal of the European Academy of Dermatology and Venereology, 25(5):554–558.

Kurugol, S., Dy, J. G., Brooks, D. H., and et al. (2011). Pilot study of semiautomated localization of the dermal/epidermal junction in reflectance confocal microscopy images of skin. Journal of Biomedical Optics, 16(3):036005.

Laine, A. and Fan, J. (1993). Texture classification by wavelet packet signatures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11):1186–1191.

Nori, S., Rius-Diaz, F., Cuevas, J., and et al. (2004). Sensitivity and specificity of reflectance-mode confocal microscopy for in vivo diagnosis of basal cell carcinoma: A multicenter study. Journal of the American Academy of Dermatology, 51(6):923 – 930. ISSN 0190-9622.

Pellacani, G., Guitera, P., Longo, C., and et al. (2007). The impact of in vivo reflectance confocal microscopy for the diagnostic accuracy of melanoma and equivocal melanocytic lesions. Journal of Investigative Dermatology, 127(12):2759–2765.

Pellacani, G., Pepe, P., Casari, A., and et al. (2014). Reflectance confocal microscopy as a second-level examination in skin oncology improves diagnostic accuracy and saves unnecessary excisions: a longitudinal prospective study. British Journal of Dermatology, online:available. DOI: 10.1111/bjd.13148

Randen, T. and Husoy, J. H. (1999). Filtering for texture classification: a comparative study. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21:291–300.

Vural, V., Fung, G., Krishnapuram, B., and et al. (2009). Using local dependencies within batches to improve large margin classifiers. Journal of Machine Learning Research, 10:183–206.

Wiltgen, M., Gerger, A., Wagner, C., and et al. (2008). Automatic identification of diagnostic significant regions in confocal laser scanning microscopy of melanocytic skin tumors. Methods of Information in Medicine, 47(1):14–25.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5:1205–1224. ISSN 1532-4435.

Table 1: Numerical results for presented DEJ delineation algorithms. Top: DEJ delineation errors, in terms of mean (μ) ± standard deviation(σ) across single stacks, as well as percentage of tiles with error <15μ*m*. The results are shown as the average across all stacks, as well as best and worst case for both dark and fair skin. Bottom: Confusion matrices, showing percent classification/misclassification rates across all 15 dark and all 15 fair skin stacks. Note that the dark skin algorithm generates two regions with one boundary while the fair skin algorithm generates three regions with two boundaries.

DARK SKIN			FAIR SKIN			
Epidermis - Dermis			Epidermis-Transition		Transition-Dermis	
Stack	$Error_{\mu\pm\sigma}$(μm)	Error<15μm	$Error_{\mu\pm\sigma}$(μm)	Error<15μm	$Error_{\mu\pm\sigma}$(μm)	Error<15μm
Average	7.9 ±6.4	71%	8.3 ±5.8	68%	7.6 ±5.6	75%
Best	3.0 ±2.6	99%	3.7 ±2.9	100%	2.2 ±1.8	100%
Worst	13.2 ±11.6	42%	20.0 ±9.7	35%	17.2 ±8.8	51%

CONFUSION MATRICES

		DARK SKIN		FAIR SKIN		
		Algorithm				
		Epidermis	Dermis	Epidermis	Transition	Dermis
Expert	Epidermis	89%	11%	64%	29%	13%
	Transition			33%	41%	26%
	Dermis	13%	87%	6%	20%	75%

FIGURE LEGENDS

Figure 1. Example images from RCM stacks of (a) dark and (b) fair skin. In the images of dark skin, basal cells (red arrows), the ring-like patterns of basal cells on dermal papillae (yellow ellipse) and enclosed papillary dermis (green arrows) can be easily distinguished. On the other hand, in fair skin, the contrast is not as strong between the ring-like patterns of basal cells (enclosed by yellow ellipses) and papillary dermis (green ellipse). The patterns can however be distinguished using textural appearance.

Figure 2. Example images from an RCM stack of 40 images (1µm depth spacing) of dark skin, showing layers from lower epidermis, DEJ and papillary dermis. The topmost row shows images in depth (axial views), from left to right, collected at 0µm (epidermis), 16µm (DEJ), and 26µm (papillary dermis), with respect to the initial imaging level at the stratum corneum-granular layer boundary. Sagittal and coronal sections, oriented perpendicular to the plane of this page and located at the red and green lines in axial views, are shown in the second and third row, respectively. In all the views, DEJ boundary drawn by expert clinicians and the algorithmic delineation are shown in light blue and orange lines, respectively. Scale bars in sagittal and coronal views show 12.5µm.

Figure 3. Example images from an RCM stack of 60 images (1µm depth spacing) of fair skin. En-face images at depths of 7µm (epidermis), 23µm (DEJ), and 51µm (papillary dermis), from initial imaging level are presented in the top two rows of the figure. In these views, dermatoglyphics appear as dark bands with epidermis on both sides (shown by yellow arrows). Coronal and sagittal sections, located at the blue and orange lines in the en-face views, are shown at the bottom rows. In all views the expert and algorithmically delineated boundaries are shown by red and yellow lines respectively. In these views, dermatoglyphics appear as cavities or dark bands in the vertical direction (shown by yellow arrows). Scale bars in sagittal and coronal views show 12.5µm.

Figure 4. Delineated DEJ represented as a 3D surface. The error between expert and algorithmic segmentation boundaries is color mapped onto the surfaces. Panel (a) shows the DEJ delineated in a dark skin stack, while panel (b) shows two boundary surfaces generated in a fair skin stack. The upper figure on the right shows the upper epidermis-to-transition region boundary in color, with the lower surface as translucent gray. The lower figure shows the lower transition-to-dermis region boundary in color, while the upper surface is translucent gray.