Existence and perception of textural information predictive of atypical nevi - preliminary insights

Paul Wighton[a,b,c] and Tim K. Lee[a,b,c] and David McLean[b,c] and Harvey Lui[b,c] and M. Stella Atkins[a]

[a]School of Computing Science, Simon Fraser University, Burnaby BC, Canada;

[b]BC Cancer Research Centre, Vancouver BC, Canada;

[c]Photomedicine Institute, Department of Dermatology and Skin Science, University of British Columbia and Vancouver Coastal Health Research Institute, Vancouver BC, Canada

ABSTRACT

Texture is known to predict atypicality in pigmented skin lesions. This paper describes an experiment that was conducted to determine 1) if this textural information is present in the center of skin lesions, and 2) how color affects the perception of this information. Images of pigmented skin lesions from three categories were shown to subjects in such a way that only textural information could be perceived; other factors known to predict atypicality were removed or held constant. These images were shown in both color and grayscale. Each subject assigned a score of atypicality to each image.

The experiment was conducted on 5 subjects of varying backgrounds, including one expert. Each subject's accuracy under each modality was measured by calculating the volume under a 3-way ROC surface. The modalities were compared using the Dorfman-Berbaum-Metz (DBM) method of ROC analysis, giving a p-value of 0.8611. Therefore the null hypothesis that there is no difference between the predictive power of the modalities cannot be rejected. Also, a two one-sided test of equivalence (TOST) was performed giving a p-value pair of < 0.01; strong evidence that the textural information is independent of color.

Additionally, the subjects' accuracies were compared to a set of random readers using the DBM and TOST methods. This was done for accuracies under the color modality, the grayscale modality and both modalities simultaneously. The results (all p-values < 0.001) confirm the existence of textural information predictive of atypia in the center of pigmented skin lesions.

Keywords: Image Perception; Observer Performance Evaluation; ROC Methodology

1. INTRODUCTION

Malignant melanoma poses a significant risk. In Canada, while the overall age-standardized mortality rate of cancer is decreasing, the mortality rate of melanoma continues to rise with a 2-3% annual increase in incidence over the last thirty years.[1] Prognosis for advanced melanoma remains poor, with a five-year survival rate of around 50% for lesions thicker than 3.5mm. However if melanoma is detected early (while the lesions thickness is less than 1.5mm), the five-year survival rate is over 90%.[2] Therefore, the early diagnosis of melanoma is critical so that it can be completely excised while it is still localized. A reliable automated diagnosis system would enable many more high-risk patients to be monitored on a regular basis. Automated diagnostic systems take as input an image of a pigmented skin lesion under high magnification. Features that are known to predict

Further author information: (Send correspondence to Paul Wighton.)
Paul Wighton: pwighton@sfu.ca
Tim K. Lee: tlee@bccrc.ca
David McLean: david.mclean@ubc.ca
Harvey Lui: harvey.lui@ubc.ca
M. Stella Atkins: stella@cs.sfu.ca

Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment,
edited by Berkman Sahiner, David J. Manning, Proc. of SPIE Vol. 6917, 69170J, (2008)
1605-7422/08/$18 · doi: 10.1117/12.770793

malignancy are then extracted from the image. Standard statistical or machine learning techniques are then used to discriminate benign and malignant lesions.

It is widely recognized that color is an important feature for the in vivo diagnosis of atypical nevi. Color is one of the four criteria included in the ABCD rule of dermatoscopy as well as Menzie's scoring method and the 7-point checklist.[3] Texture, which can be defined as the spatial variation of color, is also considered to be an important diagnostic criterion.* The asymmetry measure of color or structure in the ABCD rule could be considered in some respects to be a measure of texture. In Menzie's scoring method, 'symmetry of pattern within the lesion'[3] is also a measure of texture. Additionally, most of the positive features defined in Menzie's scoring method (blue-white veil, multiple brown dots, pseudopods, radial streaming, etc.) can be considered as specific textural patterns. The 7-point checklist can also be seen as exclusively searching for specific textural patterns.

A significant effort has been made to quantify the textural properties of skin lesions to aid in the automatic diagnosis of malignant and atypical nevi.[4–11] However, amongst dermatologists, there is no single agreed-upon list of textural patterns that are indicative of atypia. Moreover, in the signal processing and image analysis community, there is no consensus on the theoretical definition of texture.

Currently, dermatologists are more proficient at interpreting textural information than automated methods. It is hoped that some insight can be gained by observing experts. This paper describes an experiment conducted to 1) confirm the existence textural information in the center of skin lesions that is predictive of atypia and 2) determine how the presence or absence of color affects subjects' perception of this information. While the existence of textural information is not disputed, it is unclear *where* it exists. Some dermatologists may hypothesize that most of the textural information would exist at the periphery of the lesion, however this textural information would be more difficult to isolate and test. The latter objective is of great practical importance because evidence to support either claim would give strong motivation to base future models of texture upon either a single grayscale channel or a three-channel colorspace.

Section 2 reviews some previous work in the quantification of texture in pigmented skin lesions as well as the perception of texture. Section 3 describes the experimental design. A preliminary version of the experiment has been conducted on 5 subjects (one of which was a dermatologist), the results of which are discussed in section 4. Finally, section 5 concludes with a critique and a discussion of future work.

2. PREVIOUS WORK

This section discusses two largely disparate, yet relevant areas of previous work: the quantification of texture in pigmented skin lesions and the human perception of texture.

2.1 Quantification of texture in pigmented skin lesions

The analysis of textural patterns in skin lesions is an important aspect of the dermascopic diagnosis of melanoma. Many methods have been proposed to automatically quantify this textural information. Round et al. analyze the texture of fine lines present in the skin's surface. These lines are perturbed by malignant lesions moreso than benign ones. They propose a feature to measure the degree to which these skin lines are perturbed.[4] She et al. propose similar metrics to measure skin line direction as well as skin line intensity.[5] Patwardhan et al. apply an adaptive wavelet based tree structure decomposition technique, and compute statistics based on the energy of the decomposed image.[6] Deshabhoina et al. compute texture features based on second-order histogram analysis to differentiate between malignant melanoma and benign seborrheic keratosis.[7] Manousaki et al. use the intensity of the image to compute a height field. They then compute the fractal dimension and the lacunarity of this height field.[8] Lacunarity is a parameter used to differentiate fractal surfaces with similar fractal dimensions. Murali et al. use a co-occurrence matrix model to identify a specific textural pattern (dark structureless areas in skin lesions) which is predictive of melanoma.[9] Yuan et al. make use of an autoregressive filter, apply a non-linear transformation to the filter's response and then use a support vector machine with a polynomial kernel of degree 4 to differentiate between benign and malignant nevi.[10] All of these methods operate on a single grayscale channel, rather than on a 3-channel colorspace.

*Throughout this paper, we adopt this image processing definition of texture, and not the clinical concept of texture as representing surface 'roughness'.

Little work has been done on the quantification of texture in pigmented skin lesions using models operating in 3-channel colorspaces. Seidenari et al. analyzed the distance between colors in RGB colorspace at different spatial resolutions.[11] They compute several statistics based on these RGB distances (max, mean, variance) to quantify pigment distribution and use discriminant analysis to differentiate between benign, atypical and malignant lesions. They report a sensitivity and specificity of 87.5% and 85.7% respectively with this method. However, since the statistics computed are independent of spatial relationships, they likely represent color *variation* at certain spatial resolutions rather than *texture*.

2.2 Perception of texture

There is a large body of experimental research on the perception of texture.[12–16] These experiments generally task human subjects with differentiating textures based on some set of properties (such as coarseness, roughness, repetition, etc.). Mathematical metrics to mimic human perception are then formulated. These experiments have an explicitly defined task, yet are still somewhat subjective in nature. They also mostly operate on grayscale images. While the results seem promising in defining a small set of perceptual attributes (between 3 and 6) that humans use to perceive texture, the images are generally drawn from the Brodatz database,[17] or some other source in which there is a very high amount of variability. These general attributes would likely be ineffective in differentiating between textures with significantly less variance (i.e.: between types of wood, cloth, skin, etc.).

Van Rikxoort et al. studied the perception of texture images presented in both color and grayscale.[18] Human subjects were shown 180 tiles of various textures and asked to sort them into 6 categories; no further instructions were given. They report that the average consensus score did not change much when moving from color to grayscale tiles. However, they also report a relatively low average consensus score between participants. They seem skeptical of the existence of consistent human texture classification scheme. Given the subjective nature of the experiment however, this is not surprising.

To draw an analogy, suppose an experiment was conducted where various coins from various currencies were given to subjects who are tasked with sorting them into an arbitrary number of categories. Many sorting strategies may be employed (sorting based on value, size, color, country of origin, date minted, etc.) and to expect human subjects to employ the same sorting strategy given no further instructions would seem somewhat optimistic. If however, instructions were given to sort the coins based on say value, not only would consensus likely improve, but the subjects' performance could now be evaluated objectively. It is our opinion that perceptual experiments should be framed with an explicitly defined task, and whenever possible, objectively evaluated.

We therefore designed this experiment to study the effect of color on the perception of texture as it applies to a very specific and well defined task. Additionally, we evaluate performance in an objective fashion (by comparing to a 'gold standard' or ground truth). Within this greatly restricted context, we believe it is realistic to hypothesize that humans (specifically, dermatologists) have a consistent and practical classification scheme.

3. METHODS

A fully crossed study involving two modalities (color and grayscale) was conducted. Thirty images of skin lesions from three different categories were randomly selected from a dataset for which ground truth is known. A portion of each skin lesion is shown to the subject is such a way that only textural information can be perceived (see section 3.2 for details). Subjects then view and assign a confidence score to each image, representing their confidence that the lesion is atypical. The subjects' accuracies are estimated by computing a 3-way ROC surface, and calculating the volume under the ROC surface (VUS). The experiment was approved by the Simon Fraser University Office of Research Ethics.

3.1 The dataset

The data, provided by the BC Cancer Research Centre (BCCRC), consists of 246 images of skin lesions collected from between 1994 and 1998 using a video microscopy device. The camera has a fixed focal length, 20 times magnifying lens and a halogen bulb to provide consistent lighting. The images produced were in RGB format with a resolution of 512x486. The spatial resolution for each pixel is 0.033mm x 0.025mm. Each lesion is labeled as either 'clinically benign' (benign, no histopathology) 'benign' (benign, with histopathology), 'dysplastic' (atypical, with histopathology) or 'melanoma' (malignant, with histopathology)

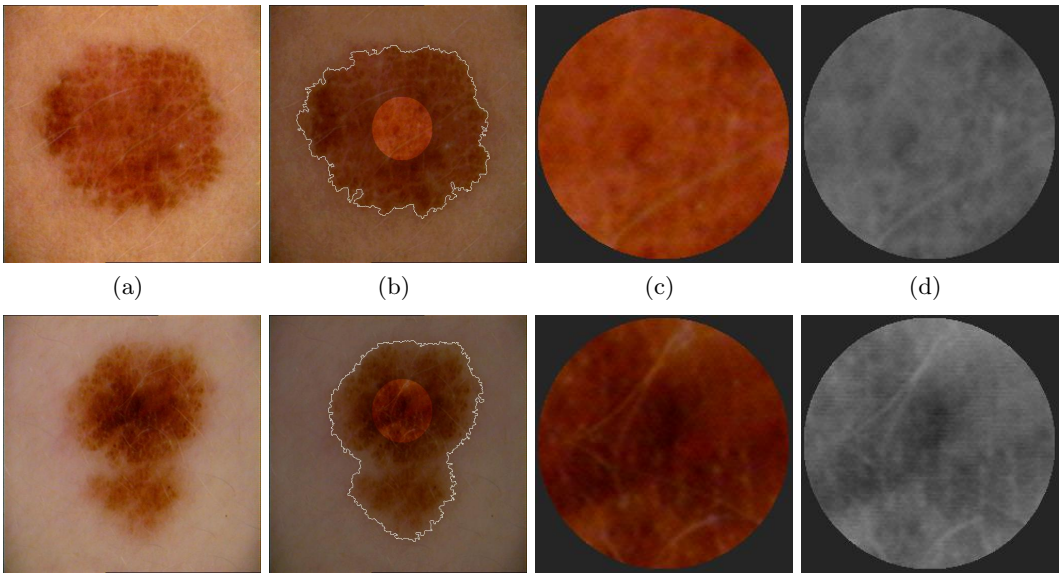

Figure 1. The two modalities under which the lesions were viewed: a), e) a dermascopic image of a benign and dysplastic nevus respectively; b), f) Selecting the region of interest for the experimental modalities; c), g) The lesions in the color modality; d), h) The lesions in the grayscale modality.

The images were first pre-processed to remove occluding hairs[19] so that the boundary between the lesion and the surrounding healthy skin could be accurately determined.[20] It is important to note that the images passed to any of the modalities are *not* pre-processed in any way; that is, all occluding hair is present in both modalities. Pre-processing was performed *only* so that the segmentation would be reasonably accurate.

After the pre-processing stage, the results were visually examined in order to ensure the segmentation was reasonable. Ten images were deemed to have an unreasonable segmentation and were removed from the dataset. Also, because of the clinical importance of identifying atypical lesions *before* they become malignant, the melanomas were also excluded from the experimental dataset.

3.2 Creating the textural images

For each image, the largest circle that can be inscribed within the lesion without intersecting the lesion border is determined. The center of this circle is used to define a circular region of interest (ROI) with a radius of 3.5mm. A test is conducted to ensure this ROI does not intersect the lesion border. Certain properties of the lesion border are known to predict atypicality;[3] therefore it is important that this border information is not present in the final textural images. The ROI uses the centre of the largest inscribable circle as opposed to the centroid because the centroid can be close to the border (or even outside the lesion entirely) on highly asymmetrical lesions. This process more accurately locates the 'core' of the skin lesion.

The image is then cropped to the bounding box of this circle. Pixels outside of the region of interest are set to a dark grey. A grayscale version of this image is then created by transforming the image to L*a*b* colorspace (using a reference whitepoint of D50) and taking the L* channel. The creation of these images is illustrated in figure 1.

3.3 The experimental procedure

To begin, each subject was given the questionnaire asking about relevant experience with pigmented skin lesions in both a clinical and research capacity. Next, they were given written instructions detailing the purpose of the experiment as well as how it will be conducted. They then reviewed online tutorials describing the Menzie's method and the 7-point checklist[21,22] for as long as they wished. Emphasis was placed on reviewing the tutorials

to develop an intuitive sense of 'benign' and 'malignancy' rather than learning the explicit details of these methods.

The experiment itself was divided into two phases: the 'training' phase and the 'data collection' phase. These phases were identical except that the training phase had fewer samples, and was intended to allow the subject to become familiar with the experimental procedure. No data gathered during the training phase was used. The training phase consisted of 6 samples. The data collection phase consisted of 10 clinically benign, 10 pathologically benign and 10 dysplastic samples. All samples were randomly selected from the dataset. For each sample, two images (color and grayscale) are generated as described in section 3.2. Since each sample is viewed twice, there exists the possibility that a sample could be recognized from a viewing under a previous modality. To minimize this ordering effect, each image was randomly rotated by either 0, 90, 180 or 270 degrees.

The subject then views each image and must rank the likelihood that the lesion is atypical on a scale of 1-100. Even though a discrete confidence rating scale is typically used (i.e. a scale from 1-5), a quasi-continuous scale (1-100) is used here. This scale is chosen because while both scales are equivalent statistically,[23] the finer resolution will likely aid in understanding how the images are interpreted. The subject then says the score out loud and it is recorded by the administrator. The subject views and scores all 60 images. No feedback on the subject's performance was given at any time during the experiment.

3.4 Analysis

This section details how the confidence scores gathered are analyzed to determine 1) how color affects the perception of this textural information and 2) if textural information predictive of atypia exists in the centre of pigmented skin lesions.

3.4.1 The effect of color

The Dorfman-Berbaum-Metz method of ROC analysis[24] was employed to determine if the accuracies of the subjects under the two modalities differed significantly. Let θ_{ij} represent the measure of accuracy of the i^{th} reader under the j^{th} modality.

The DBM method begins by computing jackknife pseudovalues for each estimate of accuracy. The pseudovalues (Y_{ijk}) estimate the contribution of each confidence score to the overall measure of accuracy and are computed by:

$$Y_{ijk} = n\theta_{ij} - (n-1)\theta_{ijk}, k = 1 \ldots n \tag{1}$$

Where θ_{ijk} is the measure of accuracy (in this case VUS) of the i^{th} reader under the j^{th} modality when the k^{th} sample is omitted and n is the number of samples.

ANOVA is then used to compare the color pseudovalues to the grayscale pseudovalues to determine if the accuracies under the two modalities differ significantly. The hypotheses can be stated as:

$$H_0 : \frac{1}{nm} \sum_{\forall(i,k)} Y_{i1k} - \frac{1}{nm} \sum_{\forall(i,k)} Y_{i2k} = 0 \tag{2}$$

$$H_1 : \frac{1}{nm} \sum_{\forall(i,k)} Y_{i1k} - \frac{1}{nm} \sum_{\forall(i,k)} Y_{i2k} \neq 0 \tag{3}$$

Where m is the number of subjects.

While the DBM method of ROC analysis can determine if the accuracies of the modalities differ significantly, it cannot determine if the accuracies are *equivalent*. Therefore a paired, two one-sided test of equivalence (TOST)[25,26] was also performed to determine if the accuracies of the modalities are equivalent. As in any statistical test of equivalence, a reasonable region of equivalence (ϵ) must be defined before the test can be conducted. The TOST procedure is the simplest form of equivalence testing; the null hypothesis is that the means are unequal:

$$H_0 : \frac{1}{nm\sigma} \sum_{\forall(i,k)} Y_{i1k} - Y_{i2k} \leq -\epsilon \vee \frac{1}{nm\sigma} \sum_{\forall(i,k)} Y_{i1k} - Y_{i2k} \geq \epsilon \tag{4}$$

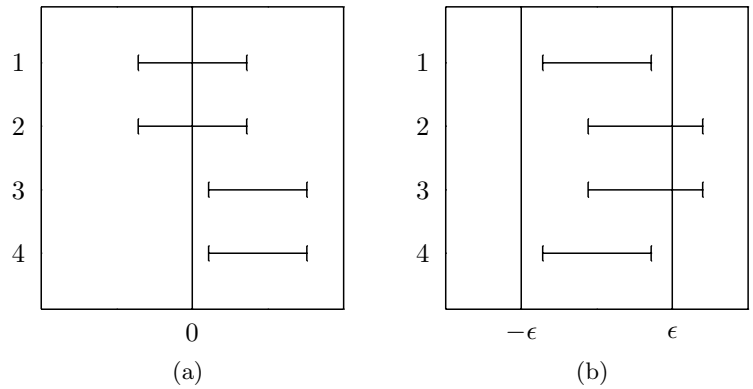

Figure 2. Confidence intervals for a) a difference test and b) an equivalence test, illustrating all four possible outcomes.

$$H_1 : -\epsilon < \frac{1}{nm\sigma} \sum_{\forall(i,k)} Y_{i1k} - Y_{i2k} < \epsilon \tag{5}$$

Where σ is the standard deviation of the pseudovalues (which is already assumed to be equal between modalities under the assumptions of ANOVA and student's t-test).

As its name implies, TOST performs an equivalence test by performing two one-sided difference tests. A rejection of both of these tests is required to reject the TOST null hypothesis. To perform a TOST at the $100(1-\alpha)\%$ confidence level, two one-sided tests are performed at the $100(1-2\alpha)\%$ confidence level. The standard choice of $\alpha = 0.05$ was used, therefore two one-sided tests at the 90% confidence level were performed.

Performing a test of equivalence in addition to a standard test of difference is a much more comprehensive way of analyzing data. Consider the four possible outcomes shown in figure 2: In case 1, the confidence interval of the difference test crosses 0, therefore the null hypothesis (that the means of the groups are equal) cannot be rejected. Under the equivalence test, the confidence interval is completely contained by the equivalence interval ($\pm\epsilon$), therefore the null hypothesis (that the means of the groups are not equivalent) can be rejected. Therefore we can conclude that there is *no evidence of a difference* in the means and that the means are *equivalent*. Similarly in case 2, we can conclude there is *no evidence of a difference* yet *no evidence of equivalence*. In case 3 we can conclude that there is a *difference* and *no evidence of equivalence*. Finally in case 4, we can conclude the means are *different* yet *equivalent*. Case 2 and 4 illustrate the potential dangers of only performing a standard statistical test of difference (t-test, ANOVA, etc.). In case 2, while no respectable statistician would claim that absence of evidence of a difference implies equality or equivalence, this logical fallacy often occurs when statistical methods are applied naïvely. Case 4, however, is much more dangerous. Here the difference test allows one to justifiably conclude that there is a statistically significant difference. However under the equivalence test, the means are equivalent. Thus while there may be a *statistically* significant difference, the difference is not *clinically* significant. Care must be taken not to mistake these two, especially when the power to detect a statistical significance is great due to a large number of observations.

All that is left is to choose an appropriate value for ϵ. Usually some clinical justification is used to determine a suitable value for ϵ, however it is unclear how a difference in VUS can be reliably transformed into a difference in pseudovalues. Therefore Wellek's recommendation for a strict interval of equivalence in terms of standard deviation ($\epsilon = \pm 0.25\sigma$) will be used.[27]

3.4.2 The existence of textural information

The methodology described in section 3.4.1 (DBM+TOST) will also be used to determine if there exists textural information in the centre of pigmented skin lesions that is predictive of atypia. If there is no information in the images, the subjects' accuracies should be no better than random guessing. Therefore if there is no information in the images, the subjects will not outperform random readers with statistical significance. To test this, 5 random readers were created and confidence scores were assigned randomly to all images. The accuracy of the

Reader	Color VUS	Grayscale VUS	Medical Training (years)	Clinical Experience (years)
Subject 1	0.484	0.387	0	0
Subject 2	0.383	0.374	0	0
Subject 3	0.337	0.297	2	0
Subject 4	0.508	0.587	9	18
Subject 5	0.282	0.301	0	0
Random	0.173	0.173	N/A	N/A

Table 1. Summary of subjects' accuracies as well as relevant experience.

random readers were calculated (θ_{ij}, $i = 6 \dots 10$) and their pseudovalues (Y_{ijk}, $i = 6 \dots 10$) were computed as in section 3.4.1. Three one-way ANOVA and TOST tests were then used to compare 1) The accuracy of the subjects reading color images vs. random readers (Y_{i1k}, $i = 1 \dots 5$ vs. Y_{i1k}, $i = 6 \dots 10$) 2) The accuracy of the subjects reading grayscale images vs. random readers (Y_{i2k}, $i = 1 \dots 5$ vs. Y_{i2k}, $i = 6 \dots 10$) and 3) The accuracy of the subjects reading both color and grayscale images vs. random readers (Y_{ijk}, $i = 1 \dots 5, j = 1, 2$ vs. Y_{ijk}, $i = 6 \dots 10, j = 1, 2$). For these TOST tests, Wellek's definition of a liberal test of equivalence ($\epsilon = \pm 0.5\sigma$) will be used.[27] Since we would ultimately like to show that the humans outperform the random readers, using a liberal definition of equivalence is more reasonable as this ultimately makes any positive result more conclusive.

4. RESULTS

The experiment was conducted, as described in section 3.3, on 5 subjects with varying degrees of medical training and experience working with pigmented skin lesions. Their accuracies as well as relevant experience is summarized in table 1.

4.1 The effect of color

The DBM+TOST method of ROC analysis as described in section 3.4.1 was employed to analyze the effect of color on the accuracies of the subjects. The ANOVA test of difference on the pseudovalues yielded a p-value of 0.8611, implying that there is no evidence to reject the null hypothesis that the accuracies are the same. A two one-sided test of equivalence was performed as described in section 3.4.1. The p-value pair for this TOST test is $(0.007, 0.002)$, allowing the rejection of the null hypothesis that the accuracies are not equivalent. Therefore one can conclude that no textural information predictive of atypia is lost when viewing dermascopic images in grayscale.

4.2 The existence of textural information

Five automated random readers assigned confidence scores to all images. Their accuracies were computed and the DBM+TOST method was used to compare the accuracies of the subjects to the accuracies of the random readers. Comparing the accuracies in the color modality, the ANOVA test of difference yields a p-value of 6.9×10^{-4}; the TOST test of equivalence yields a p-value pair of $(0.00, 0.16)$. Similarly, comparing the accuracies in the grayscale modality, the p-value of the ANOVA and p-value pair of the TOST test is 1.4×10^{-5} and $(0.00, 0.48)$ respectively. Finally, comparing the accuracies of both modalities simultaneously yields a p-value of 4.4×10^{-8} for the ANOVA test and a TOST p-value pair of $(0.00, 0.22)$.

In all three cases, the difference test is statistically significant, *and* there is no evidence to reject the null hypothesis of the equivalence test (that the accuracies are not equivalent). Therefore we can conclude that the accuracies of the subjects is greater than that of the random readers, and that textural information predictive of atypia exists within the center of skin lesions.

5. CONCLUSIONS AND DISCUSSION

We have shown that textural information exists in the centre of skin lesions that predicts atypia, and the perception of this textural information is independent of color. While these initial results are encouraging, our work is far from done. This section offers a critique of the experiment performed, and discusses avenues for future work.

5.1 Critique

There are two major limitations to this experiment. First, only one of the five subjects was a dermatologist. While the use of lay people further reinforces the claim of the existence of textural information (because it is expected that dermatologists would outperform lay people), it weakens the claim that perception of textural information is independent of color. We are very much interested in how *dermatologists*, not lay people perceive this information. One should exercise caution in concluding that *dermatologists'* perception of this textural information is independent of color. However these promising results motivate a second study in which all subjects are dermatologists.

Second, it is known that variance is a predictor of atypia in skin lesions.[3] It is possible that the subjects were perceiving variance and not texture. Recall that texture is the *spatial* relation of color or intensity; variance is independent of spatial relationships. One method of accounting for this possibility is to create an automated reader that assigns confidence scores based on the image variance. If the subjects can outperform this automated reader, then this would strengthen then evidence of the existence of textural information. However conversely, if the subjects do not outperform this variance-based automated reader, this does not prove non-existence of textural information. This test was conducted, but statistical significance was not achieved. Individually, however, the dermatologist nearly obtained statistical significance ($p=0.0569$ under the grayscale modality). It is hypothesized that a second experiment, one in which all subjects are dermatologists, would be able to outperform this automated reader with statistical significance. Another method for accounting for the possibility that subjects are perceiving variance would be to equalize the variance of the images. This will also be considered in the design of future experiments.

5.2 Future Work

Future work includes conducting this experiment with all subjects are dermatologists to address the concerns noted in section 5.1. Also an experiment where all the subjects were dermatologists would allow us to test the degree of consistency to which dermatologists perceive this texture. If this is shown to be sufficiently consistent, then a model observer can be constructed to approximate the dermatologist's perception. This model can then be validated on a new dataset by comparing it to dermatologists. Finally, given a validated model observer for dermatologists' perception of texture, a reliable quantification can be proposed for inclusion into automated diagnostic systems.

ACKNOWLEDGMENTS

This work is supported in part by a discovery grant from the Natural Sciences and Engineering Research Council of Canada, a Canadian Dermatology Foundation grant and the CIHR Skin Research Training Centre. The authors would also like to thank all participants who volunteered their time.

REFERENCES

1. Canadian Cancer Society/National Cancer Institute of Canada, "Canadian cancer statistics 2007." Toronto, Canada, 2007.
2. M. Lens and M. Dawes, "Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma," *British Journal of Dermatology* **150**(2), pp. 179–185, 2004.
3. R. H. Johr, "Dermoscopy: alternative melanocytic algorithms the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist," *Clinics in Dermatology* **20**, pp. 240–247, 2002.
4. A. J. Round, A. W. G. Duller, and P. J. Fish, "Lesion classification using skin patterning," *Skin Research and Technology* **6**, pp. 183–192(10), November 2000.

5. Z. She, Y. Lui, and A. Damatoa, "Combination of features from skin pattern and ABCD analysis for lesion classification," *Skin Research and Technology* **13**, pp. 25–33(9), February 2007.
6. S. V. Patwardhan, A. P. Dhawan, and P. A. Relue, "Classification of melanoma using tree structured wavelet transforms," *Computer Methods and Programs in Biomedicine* **72**, pp. 223–239(17), November 2003.
7. S. V. Deshabhoina, S. E. Umbaugh, W. V. Stoecker, R. H. Moss, and S. K. Srinivasan, "Melanoma and seborrheic keratosis differentiation using texture features," *Skin Research and Technology* **9**(4), pp. 348–356, 2003.
8. A. G. Manousaki, A. G. Manios, E. I. Tsompanaki, and A. D. Tosca, "Use of color texture in determining the nature of melanocytic skin lesions - a qualitative and quantitative approach," *Computers in Biology and Medicine* **36**(4), pp. 419–427, April 2006.
9. A. Murali, W. V. Stoecker, and R. H. Moss, "Detection of solid pigment in dermatoscopy images using texture analysis," *Skin Research and Technology* **6**, pp. 193–198(6), November 2000.
10. N. Xiaojing Yuan; Zhenyu Yang; Zouridakis, G.; Mullani, "Svm-based texture classification and application to early melanoma detection," *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE* , pp. 4775–4778, August 2006.
11. S. Seidenari, G. Pellacani, and C. Grana, "Pigment distribution in melanocytic lesion images: a digital parameter to be employed for computer-aided diagnosis," *Skin Research and Technology* **11**, pp. 236–241(6), November 2005.
12. H. Tamura, M. S., and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man and Cybernetics* **8**(6), pp. 460–73, 1978.
13. B. Julesz and J. R. Bergen, "Textons, the fundamental elements in preattentive vision and perception of textures," *Bell Systems Technical Journal* **62**(6), pp. 1619–45, 1983.
14. J. Malik and P. Perona, "A computational model of texture perception," tech. rep., Berkeley, CA, USA, 1989.
15. A. Rao and G. Lohse, "Identifying high level features of texture perception," *Graphical Models and Image Processing* **55**(3), pp. 218–33, 1993.
16. R. Cho, V. Yang, and P. Hallett, "Reliability and dimensionality of judgments," *Perception and Psychophysics* **62**(4), pp. 735–52, 2000.
17. P. Brodatz, *Textures: A photographic Album for Artists and Designers*, Dover, 1966.
18. E. M. van Rikxoort, E. L. van den Broek, and T. E. Schouten, "Mimicking human texture classification," *Human Vision and Electronic Imaging X* **5666**(1), pp. 215–226, SPIE, 2005.
19. T. Lee, V. Ng, R. Gallacher, A. Coldman, and D. McLean, "Dullrazor(r): A software approach to hair removal from images," *Computers in Biology and Medicine* **27**, pp. 533–543(11), November 1997.
20. M. A. King, T. K. Lee, M. S. Atkins, and D. I. McLean, "Automatic nevi segmentation using adaptive mean shift filters and feature analysis," in *Medical Imaging 2004: Image Processing. Edited by Fitzpatrick, J. Michael; Sonka, Milan. Proceedings of the SPIE, Volume 5370, pp. 1730-1737 (2004).*, 2004.
21. Dermoscopy.org, "Menzies' method for the diagnosis of melanoma (tutorial)," 2007. [Online; accessed 2007/12/04. http://www.dermoscopy.org/consensus/2c.asp].
22. Dermoscopy.org, "7-point checklist (tutorial)," 2007. [Online; accessed 2007/12/04. http://www.dermoscopy.org/consensus/2d.asp].
23. H. E. Rockette, D. Gur, and C. Metz, "The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques.," *Investigative Radiology February* **27**, pp. 169–172, 1992.
24. D. Dorfman, K. Berbaum, and C. Metz, "Receiver operating characteristic rating analysis. generalization to the population of readers and patients with the jackknife method," *Investigative Radiology September* **27**, pp. 723–31, 1992.
25. D. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.," *Journal of pharmacokinetics and biopharmaceutics* **15**, pp. 657–80, dec 1987.

26. L. E. Baker, E. T. Luman, M. M. McCauley, and S. Y. Chu, "Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage," *American Journal of Epidemiology* **156**, pp. 1056–61, December 2002.
27. S. Wellek, *Testing Statistical Hypotheses of Equivalence*, CRC Press LLC, 2003.