MELANOMA CLASSIFICATION FROM HIDDEN MARKOV TREE FEATURES

Marco F. Duarte,[m] *Thomas E. Matthews,*[d] *Warren S. Warren,*[d] *Robert Calderbank* [d]

[m] University of Massachusetts, Amherst, MA 01003
[d] Duke University, Durham, NC 27707

ABSTRACT

Melanoma detection relies on visual inspection of skin samples under the microscope via a qualitative set of indicators, causing large discordance among pathologists. New developments in pump-probe imaging enable the extraction of melanin intensity levels from skin samples and provide baseline qualitative figures for melanoma detection and classification. However, such basic figures do not capture the diverse types of cellular structure that distinguish different stages of melanoma. In this paper, we propose an initial approach for feature extraction for classification purposes via Hidden Markov Tree models trained on skin sample melanin intensity images. Our experimental results show that the proposed features provide a *mathematical microscope* that is able to better discriminate cellular structure, enabling successful classification of skin samples that are mislabeled when the baseline melanin intensity qualitative figures are used.

***Index Terms*—** Image processing, wavelet transform, hidden Markov tree, melanoma detection and classification.

1. INTRODUCTION

Melanomas are among the most commonly occurring cancers, but they are clinically challenging to diagnose. For example, from 1991 to 2005, while overall cancer death rates decreased by 19% in men and 11% in women, the death rate for melanoma in the United States increased more than 5% [1]. Early detection remains difficult but is critical for successful treatment; five year survival rates fall from 98% for local cancers to 16% for metastatic melanoma [2].

The current standard for diagnosis remains biopsy and histopathology, but this too results in discordant conclusions because there is no one histological criterion for melanoma — diagnosis must instead be made by subjectively weighing a series of separate indicators that may be present in atypical lesions as well. A recent study found a discordance rate of 14% among pathologists for melanoma diagnosis [3]. Doctors must therefore err on the side of caution, which leads to an excess of false positive diagnoses and increased medical costs and emotional trauma from unnecessary surgeries, lymph node biopsies, and other treatments.

Melanoma presents a promising target for optical diagnosis both because suspicious lesions are accessible and disease occurs within a few hundred micrometers of the skin surface. Since melanin carries information about the metabolism and location of melanocytes (melanin producing cells) in tissue, the distribution of eumelanin and pheomelanin, the two dominant types of melanin, could act as a marker for disease.

It was recently demonstrated that two-color pump-probe imaging of skin samples can separate eumelanin and pheomelanin, two chromophores that are key to separate melanomas from benign nevi in a highly sensitive manner [4]. Initial approaches to melanoma detection and classification based on pump-probe imaging rely on pixel-averaged eumelanin-to-pheomelanin concentration ratios, achieving remarkable performance levels [5]. However, such pixel-wise features cannot capture the morphological features that distinguish different stages of melanoma, which are evident by inspection of the obtained images (see Fig. 1).

In this paper, we leverage prior work on modeling of natural images to perform feature extraction from pump-probe skin sample images for the purpose of melanoma classification. We focus on statistical models of image wavelet coefficients that capture the scale, predominance, and intensity of the spatial features present in the image. We show empirically that the use of these statistical morphological features for classification leverages the distinct cellular features of the different stages of melanoma to improve the performance of melanoma classification over pixel-averaged methods. In a sense, wavelet analysis provides a *mathematical microscope* that tabulates features relevant to melanoma detection and classification.

2. BACKGROUND

Melanoma classification: We briefly describe several classes of melanoma according to the histological and chemical features evident in the skin samples, as illustrated in Figure 1.

Benign nevi exhibit a proliferation of small melanocytes arranged as single cells and/or nests usually distributed along the basal layer, the deepest layer in the epidermis. These samples usually contain no melanocytes in the superficial layers of the epidermis, indicating no pagetoid (outward) spread.

R. Calderbank and M. F. Duarte were supported in part by AFOSR under grant FA9550-09-1-0422, by NSF under Grant DMS-0914892, and by ONR under Grant N00014-08-1-1110. M. F. Duarte was supported in part by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch. Email: mduarte@ecs.umass.edu, {thomas.e.matthews,robert.calderbank,warren.warren}@duke.edu.

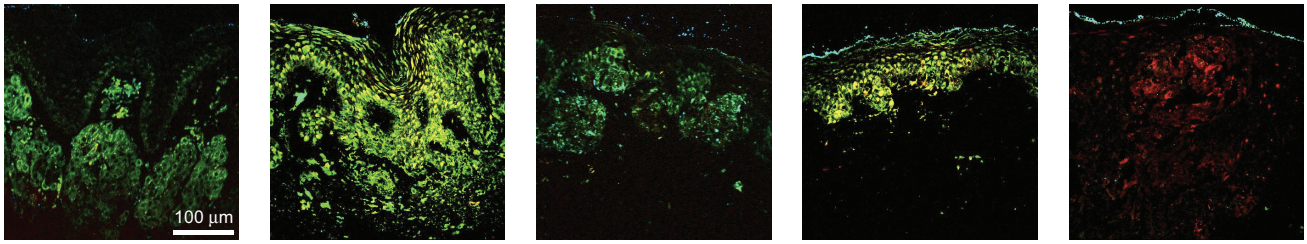

Fig. 1. *Example pump-probe images for different classes of skin samples. Orange denotes eumelanin, green denotes pheomelanin, and blue denotes surgical ink. From left: benign nevus, compound nevus, dysplastic nevus, melanoma* in-situ, *and invasive melanoma.*

Compound nevi exhibit small matured melanocytes that involve the epidermis as well as the dermis. The intra-epidermal component of these lesions is organized as single cells and nests without confluent growth pattern or presence of upper migration of melanocytes. *Dysplastic nevi* feature junctional nests of melanocytes that appear as a clustering of pigmented cells at the basal layer. Pigmented keratinocytes are often seen in the upper epidermis, but there is no pagetoid spread. *Melanoma in-situ* samples correspond to very early forms of melanoma, where the melanocytes have proliferated only radially within the base of the epidermis. In contrast, *invasive melanoma* features both radial and vertical proliferation of melanocytes. Such samples also tend to have large structure across the image, including pigmentation in the dermis.

In terms of chemical features, a large quantity of pheomelanin is often found in melanocytic benign and compound nevi, with a shift to eumelanin dominance in melanoma. Because eumelanin is photoprotective and has antioxidant properties, whereas pheomelanin can act as a photosensitizer, it has been postulated that elevated amounts of pheomelanin would lead to increased damage from ultraviolet radiation and an increased risk of malignant transformation. Dysplastic nevi display atypical growth and seem to have increased pheomelanin content compared to normal skin and other melanocytic nevi [6]. However, the chemical identity of melanin in melanomas is less clear, and some evidence exists to show that eumelanin may in fact occur in increased concentration [7]. This indicates a more heterogeneous chemical signature being characteristic of melanomas in contrast to other types of lesions. The generalizations we make are drawn from only one new study (our STM paper) and in limited sample set, and that the literature does not paint a clear or definitive picture either.

Bulk analysis of the eumelanin content alone allows for the rejection of many false positive diagnoses. We calculate a weighted average of eumelanin content across the entire image by normalizing by total melanin content in a pixel-wase fashion. Regions containing surgical ink were not considered. As shown in Fig. 2, if only raw melanin content is considered, a threshold of 38% eumelanin captured all invasive melanomas and most of the melanomas *in situ* while excluding $> 75\%$ of the dysplastic nevi. Although the eumelanin to pheomelanin ratio is not sufficient to diagnose melanoma, it may greatly improve diagnostic accuracy in conjunction with

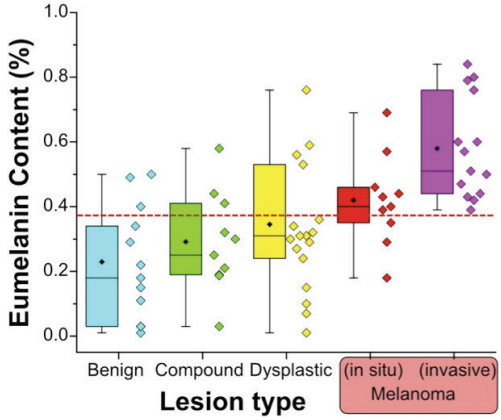

Fig. 2. *Distribution of normalized average eumelanin content of skin samples grouped by class. Overlaid on the box and whisker plot are the actual data points for the different samples. The dashed line shows the 38% average eumelanin threshold used to separate melanomas from nevi.*

complementary diagnostic techniques.

Hidden Markov trees: A widely used sparse representation in signal and image processing is the wavelet transform. The wavelet transform of an image provides a multiscale time-frequency analysis of the image content, effectively encoding the locations and scales of the image features in a compact fashion. This energy compaction property is the main reason behind the popularity of wavelet transforms for image processing and compression, including the state-of-the-art JPEG2000 standard.

In a typical 2D real-valued wavelet transform of an $\sqrt{N} \times \sqrt{N}$-pixel image $x \in \mathbb{R}^N$, each *wavelet coefficient* $w_{o,s,i,j}$ is labeled by a scale $s \in \{1, \ldots, S := \log_2(N)/2\}$, orientation $o \in \{H, V, D\}$ for horizontal, vertical, and diagonal, respectively, and offset (i, j), $1 \leq i, j \leq 2^{s-1}$. Additionally, a scaling coefficient w_0 captures the remaining energy of the signal. The image x can then be written as

$$x = w_0 \varphi + \sum_{o \in \{H,V,D\}} \sum_{s=1}^{S} \sum_{i,j=1}^{2^{s-1}} w_{o,s,i,j} \psi_{o,s,i,j},$$

where φ denotes the scaling function and $\psi_{o,s,i,j}$ denotes the mother wavelet function ψ_o for orientation o dilated to scale s and translated to offset (i, j). For convenience, we also index the wavelet coefficients and wavelet functions as $\{w_0, w_1, \ldots, w_{N-1}\}$ and $\{\varphi, \psi_1, \ldots, \psi_{N-1}\}$ using an arbitrary ordering, e.g., lexicographic.

A coefficient $w_{o,s,i,j}$ at scale s describes a portion of the signal of size $O(4^{-s})$. With 4^{s-1} such coefficients at each scale and orientation, a quad-tree provides a natural organization for the coefficients. Each coefficient at scale $s < \log_2(N)/2$ has 4 *children* at scale $s+1$, and each coefficient at scale $s > 1$ has one *parent* at scale $s-1$.

A large wavelet coefficient (in magnitude) generally indicates the presence of a singularity inside its support; a small wavelet coefficient generally indicates a smooth region. This energy compaction property causes wavelet coefficients to have a peaky non-Gaussian distribution. Thanks to the nesting of child wavelets inside their parents, edges in general manifest themselves in the wavelet domain as chains of large coefficients propagating across scales in the wavelet quad tree; we call this phenomenon the *persistence property*.

Hidden Markov Trees (HMTs) [8] offer one modeling framework that succinctly and accurately captures this joint structure in natural images. In an HMT, wavelet coefficients are modeled probabilistically using a mixture of Gaussians: one component features a large variance that models large nonzero coefficients and receives a small weight (to encourage few such coefficients), while a second component features a small variance that models small and zero-valued coefficients and receives a large weight. We distinguish these two components by associating to each wavelet coefficient w_n an unobserved hidden state $\mathcal{S}_n \in \{\mathsf{S}, \mathsf{L}\}$; the value of $\mathcal{S}_n$ determines which of the two components of the mixture model is used to generate w_n:

$$f(w_n|\mathcal{S}_n = \mathsf{S}) = \mathcal{N}(0, \sigma^2_{\mathsf{S},n}), f(w_n|\mathcal{S}_n = \mathsf{L}) = \mathcal{N}(0, \sigma^2_{\mathsf{L},n}),$$

with $\sigma^2_{\mathsf{L},n} > \sigma^2_{\mathsf{S},n}$. To generate the mixture, we apply a probability distribution to the available states: $p(\mathcal{S}_n = \mathsf{S}) = p^{\mathsf{S}}_n$ and $p(\mathcal{S}_n = \mathsf{L}) = p^{\mathsf{L}}_n$, with $p^{\mathsf{S}}_n + p^{\mathsf{L}}_n = 1$.

To simplify the model, the coefficient-dependent parameters are made equal for all coefficients within a scale; that is, the new model has parameters $\Theta = \{p^{\mathsf{L}}_s, p^{\mathsf{S}}_s, \sigma_{\mathsf{L},s}, \sigma_{\mathsf{S},s}\}_{1<s\leq S}$. We can obtain estimates of all these parameters for a set of coefficients w using maximum likelihood estimation via an expectation-maximization (EM) algorithm [8].

3. FEATURE EXTRACTION

We design a feature set for a skin sample based on the parameters of an HMT trained over its pump-probe image. The goal of feature selection is to capture the scale and orientation of the dominant features in an image, given that different types of melanoma exhibit different scales and orientations for their cellular structure. This is in contrast to the features in [5], which do not consider orientation and scale, but rather global average melanin concentration. Therefore, we select as features the likelihoods of the small state p^{S}_s for each orientation $\{H, V, D\}$ and scale $s = 1, \ldots, S$. If strong features are present in an image at a given scale and orientation, the corresponding wavelet coefficients will be large, yielding a small probability of the small state for that scale and orientation during HMT training. We perform this feature extraction on three different types of processed pump-probe images.

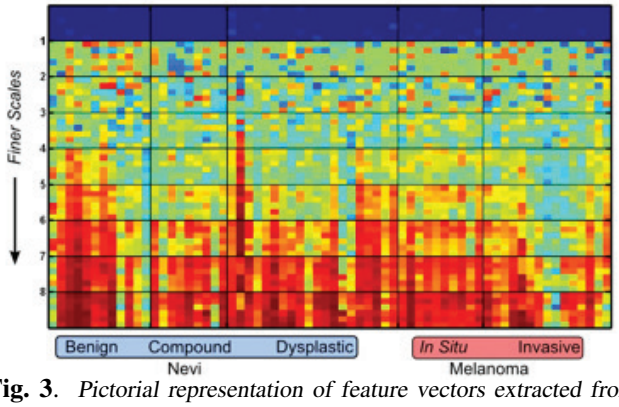

Fig. 3. *Pictorial representation of feature vectors extracted from skin samples. Each column corresponds to a different sample from the labeled classes and each row corresponds to the likelihood of a small state for a given scale. Red (dark) color denotes largest likelihood, while blue (light) colors denote smallest likelihood.*

F1: We train an HMT on a sample image depicting only total melanin concentration, i.e., we sum the eumelanin and pheomelanin concentration values for each individual pixel. While this modality gives smaller feature vectors, it does not discriminate between the presence of both types of melanin, a key distinguishing feature between classes as shown in Fig. 1.

F2: We train two HMTs individually on the eumelanin and pheomelanin concentration images. However, F2 fails to discriminate between chemically homogeneous and heterogeneous regions, since each feature relies only on the concentration of a single type of melanin.

F3: We create new images that express the percentage concentration of each type of melanin and train HMTs on each of them. In other words, the pixel of the processed images contain the ratio between the intensity of the eumelanin and pheomelanin concentrations, respectively, and their sum. F3 expresses information about chemical homogeneity.

Figure 3 shows the extracted feature vectors from the eumelanin and pheomelanin intensity images for the second type of extracted features. Each 512×512-pixel skin sample image is reduced to a feature vector containing the small state likelihoods for the two channels (one for the first type of features), three orientations, and $S = 9$ different scales. This results in a feature vector of dimension $F = 54$ (or 27, respectively). The figure shows a few qualitative differences. For nevi, small coefficients are more likely at all size scales. Melanomas are more likely to have large coefficients. However, these differences are subtle and variable, and we need sophisticated classifiers to make more rigorous statements.

4. MELANOMA CLASSIFICATION PERFORMANCE

Once the images of lesions have been reduced to an appropriate mathematical representation, they need to be classified. Support vector machines (SVMs) have been used to tackle diverse classification and regression problems and are one of the most effective tools for these tasks. To develop a clas-

Test	P_C	P_D	$1 - P_{FA}$
Melanoma vs. nevi	73%	72%	74%
Melanoma vs. nevi + SKs	61%	62%	60%
Invasive melanoma vs. nevi	57%	54%	57%
In situ melanoma vs. nevi	72%	73%	72%
Melanoma vs. benign	59%	60%	58%
Melanoma vs. dysplastic	56%	52%	60%

Table 1. *Skin sample classification results for F2 features, scales 4–9, $\sigma = 0.2$. The table shows the probability of correct classification P_C, sensitivity (probability of detection) P_D, and specificity (complement of false alarm probability) $1 - P_{FA}$ for each of the classification problems listed.*

sifier for the feature vectors of wavelet transformed images of pigmented lesions, we used a Neyman-Pearson SVM, also known as a 2ν-SVM [9].[1] This algorithm allows assignments of different weights ν_1 and ν_2 to false positives and false negatives during training, respectively. The script searches for best performing parameters ν_1 and ν_2, as well as the best bandwidth parameter σ of the Gaussian radial basis kernel function via line search. The parameter selection was performed using leave one out cross validation.

We tested several classification problems as listed in Table 1. Melanoma includes both invasive and *in situ* unless otherwise noted. Nevi includes benign, compound, and dysplastic nevi unless otherwise noted. For each binary classification problem, we tested all three feature types F1–F3. We used the 2ν-SVM classifier with radial basis function parameter values σ varying between 0.1 and 0.5. We also did feature subset selection by dropping the features corresponding to the coarsest scales, varying from $s = 1$ to $s = 4$.

Classification rates for the various problems varied from 40% to 75%, depending on the test performed and feature used. Invasive melanoma vs. nevi goes up to 65% classification, 68% sensitivity and 64% specificity when F2 features are considered instead of F1. The improvement also appears for the melanoma (both) vs. dysplastic problem, which has 68% success, 75% sensitivity and 60% specificity.

The success of wavelet scales 4–9 was comparable or even superior to that of the full feature vector for all trials. This improvement indicates that coarse features such as the distribution of melanin in the epidermis vs. the dermis are not contributing to correct classification. Instead, this analysis appears to be highlighting differences on the cellular scale. The wavelet scale $s = 4$ corresponds to features $\sim 20\mu m$ in size and smaller, which is 1–2 cells. Differences on this scale were not noticed by our previous visual inspection of these images, and the wavelet transform method is most likely provided a complementary analysis and not picking up on the same features discussed in Section 2.

This complementary information should therefore produce a different set of errors and correct false positives and negatives than bulk eumelanin classification. Indeed, we find this to be the case. We found many different classifiers which correct 100% of false negatives and 80–100% of false positives from bulk eumelanin classification. These classifiers operated on the invasive melanoma vs. nevi and melanoma *in situ* vs. nevi problems. Only the F3 features (and not F2 features) yielded classifiers which corrected 8+ of the false positives. For invasive melanoma vs. all nevi, the $\sigma = 0.2$ classification of scales 4–9 had 65% correct classification rate, 68% sensitivity and 64% while correctly calling 9 out of 10 false positives from average eumelanin classification (none of the false negatives from average eumelanin classification were invasive melanomas). For melanoma *in situ* vs. all the nevi, many different classifiers corrected all of the false negatives. The $\sigma = 0.1$ classification of scales 4–9 corrected all of the false negatives and 10 out of 10 false positives with 59% sensitivity and 56% specificity. Other conditions for these two groups gave higher sensitivity (up to 68%) and specificity (up to 64%) while correcting the false results.

The classifiers had greater success in many trials when only one category of melanoma was considered, which indicates differences between the way these two types of melanoma present. In this case we expect more subtle changes to appear between the two classifications of melanoma, especially on the cellular level. For example, Fig. 1 shows the lack of dermal pigmentation in the melanoma *in situ*. These subtle changes may additionally include enlarged nucleus, local heterogeneity of pigmentation or other features not yet discovered to which the wavelet transform is sensitive.

[1] We thank Mark Davenport for sharing the 2ν-SVM classifier code.

5. REFERENCES

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA, Cancer J. Clinicians*, vol. 59, no. 4, pp. 225–249, May 2009.

[2] "Cancer facts and figures," Tech. Rep., American Cancer Society, Atlanta, GA, 2009.

[3] B. A. Shoo, R. W. Sagebiel, and M. Kashani-Sabet, "Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center," *J. Amer. Acad. Dermatology*, vol. 62, no. 5, pp. 751–756, May 2010.

[4] D. Fu, T. Ye, T. E. Matthews, G. Yurtsever, and W. S. Warren, "Two-color, two-photon, and excited-state absorption microscopy," *J. Biomedical Optics*, vol. 12, no. 5, 2007.

[5] T. E. Matthews, I. R. Piletic, M. A. Selim, M. J. Simpson, and W. S. Warren, "Pump-probe imaging differentiates melanoma from melanocytic nevi," *Science Translational Medicine*, vol. 3, no. 71, Feb. 2011.

[6] T. G. Salopek, K. Yamada, S. Ito, and K. Jimbow, "Dysplastic melanocytic nevi contain high levels of pheomelanin: Quantitative comparison of pheomelanin/eumelanin levels between normal skin, common nevi, and dysplastic nevi," *Pigment Cell Res.*, vol. 4, pp. 172–179, 1991.

[7] H. Hara, N. Walsh, K. Yamada, and K. Jimbow, "High plasma level of a eumelanin precursor, 6-hydroxy-5-methoxyindole-2-carboxylic acid as a prognostic marker for malignant melanoma," *J. Investigative Dermatology*, vol. 102, pp. 501–505, 1994.

[8] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[9] M. A. Davenport, R. G. Baraniuk, and C. D. Scott, "Controlling false alarms with support vector machines," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006, vol. V, pp. 589–592.