

High-Level Intuitive Features (HLIFs) for Intuitive Skin Lesion Description

Robert Amelard*, *Student Member, IEEE*, Jeffrey Glaister, *Student Member, IEEE*, Alexander Wong, *Member, IEEE*, and David A. Clausi, *Senior Member, IEEE*

Abstract—A set of high-level intuitive features (HLIFs) is proposed to quantitatively describe melanoma in standard camera images. Melanoma is the deadliest form of skin cancer. With rising incidence rates and subjectivity in current clinical detection methods, there is a need for melanoma decision support systems. Feature extraction is a critical step in melanoma decision support systems. Existing feature sets for analyzing standard camera images are comprised of low-level features, which exist in high-dimensional feature spaces and limit the system's ability to convey intuitive diagnostic rationale. The proposed HLIFs were designed to model the ABCD criteria commonly used by dermatologists such that each HLIF represents a human-observable characteristic. As such, intuitive diagnostic rationale can be conveyed to the user. Experimental results show that concatenating the proposed HLIFs with a full low-level feature set increased classification accuracy, and that HLIFs were able to separate the data better than low-level features with statistical significance. An example of a graphical interface for providing intuitive rationale is given.

Index Terms—Decision support, feature extraction, melanoma, pigmented skin lesion.

I. INTRODUCTION

CUTANEOUS melanoma (i.e., melanoma of the skin) is the deadliest form of skin cancer [1]. The World Health Organization (WHO) estimated that approximately 65 000 global deaths related to melanoma occurred in the year 2000 [2]. This death toll is increasing; melanoma incidence rates have been increasing on average by 2.6% each year over the last ten years in the U.S. [3]. If caught early when the cancer is localized, a simple excision of the cancerous tissue results in a 98% five-year survival rate [4]. However, if identified late when the cancer has spread remotely, the prognosis is a bleak 15% five-year survival rate [4].

Manuscript received July 11, 2014; revised October 17, 2014; accepted October 18, 2014. Date of publication October 28, 2014; date of current version February 16, 2015. This work was supported in part by Agfa Healthcare Inc., Ontario Ministry of Economic Development and Innovation, Ontario Centres of Excellence, the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chairs program. Asterisk indicates corresponding author.

*R. Amelard is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada (e-mail: ramelard@uwaterloo.ca).

J. Glaister was with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada. He is now with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: jglaist1@jhu.edu).

A. Wong and D. A. Clausi are with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada (e-mail: alexander.wong@uwaterloo.ca; dclausi@uwaterloo.ca).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TBME.2014.2365518

Cutaneous melanoma is the cancerous growth of melanocytes, cells found at the bottom of the epidermal layer of the skin, which are responsible for producing the UV-absorbing pigment melanin. In North America, initial melanoma detection is usually done visually by a general practitioner, followed by a follow-up appointment with a dermatologist for further visual inspection. This process is time- and cost-inefficient, especially with increasing incidence rates [3]. Additionally, the following two factors make it difficult to visually identify melanoma: 1) melanoma can be very similar in appearance to benign nevi (i.e., noncancerous “moles”) at the surface during its early to midstages and 2) melanoma can take on widely varying shapes and forms.

Dermatologists commonly use metrics such as the ABCD (asymmetry, border irregularity, color patterns, and diameter) criteria [5], [6] or the seven-point checklist [7]. However, usage of these metrics is very subjective, leading to large interobserver variability [8]. Systematic objective decision support systems can help meet the demand of the rising rate of melanoma and help reduce subjectivity.

A critical step in computer-aided melanoma detection involves extracting quantitative features from images of lesions. Many existing feature extraction methods have focused on modeling the ABCD criteria using dermoscopic data (i.e., images obtained with a dermatoscope). Dermatoscopes are optical devices that manipulate light characteristics to elucidate subsurface information. Reviews of existing features can be found in [9] and [10]. Unfortunately, the clinical use of dermatoscopes is limited in North America, with a recent survey reporting less than 50% utilization in the USA [11]. We, therefore, turned to analyzing images obtained with standard consumer-grade cameras. Some feature sets have been proposed for images obtained with standard cameras (e.g., [12]–[14]), however, these feature sets combined many low-level features (LLFs) to try to approximate ABCD. The importance of high-level over low-level features has been recently discussed [15]. LLFs are (usually simple) features that were not designed to model a high-level characteristic (e.g., asymmetry). This limits the system's ability to present diagnostic rationale, which is important for user-system trust [16]. System credibility has received attention in human-computer interaction research [17]; however, these ideas have not been explicitly introduced to melanoma decision support system research.

The main contribution of this study is a set of high-level intuitive features (HLIFs) for analyzing skin lesions. HLIFs are designed explicitly to model human-observable characteristics. As such, an HLIF's design is usually more complex than that of an LLF. A decision support system that extracts HLIFs can provide intuitive diagnostic rationale to the user according to

what they would expect to observe, with the aim of increasing user-system trust. Experimental results show that concatenating a small set of HLIFs and a set of LLFs increased classification accuracy over the LLF alone. This study builds on previous work [18]–[20] by extending the previously proposed asymmetry and border irregularity HLIFs, and by proposing six new HLIFs for color variation.

The remainder of this paper is organized as follows. Section II provides a framework for designing HLIFs. Section III presents a set of ten HLIFs, which models the ABCD melanoma criteria. Section IV presents statistical analyses of the proposed HLIFs as well as experimental classification results of the HLIF with a recent full LLF set using the public databases Dermatology Information System [21] and DermQuest [22]. Results and limitations of the system are discussed in Section V and conclusions are drawn in Section VI.

II. HIGH-LEVEL INTUITIVE FEATURES (HLIFs)

This section presents HLIFs as a feature extraction framework for intuitive classification problems. Advantages of designing HLIFs are discussed, followed by general instructions for designing an HLIF. This framework is used in Section III for extracting features relevant to skin cancer detection.

A. Rationale

The success of a decision support system is highly reliant on the efficacy of feature extraction. “Good” features are specific characteristics about an image that project the data into a space where the inherent classes are well separated. However, effective feature extraction is not a trivial task.

B. LLFs

The LLFs are features that are not designed to model a high-level characteristic (e.g., color asymmetry). Many feature sets combine several LLFs to capture some high-level characteristic of an object. For example, to describe a lesion’s color, Ganster *et al.* combined simple calculations such as minimum, maximum, average, and variance of intensity and hue color channels [23]. These features do not individually model lesion color, and are not aimed at describing specific observable characteristics. This combination results in a high-dimensional feature space, leading to theoretical and computational complexities.

A benefit of using LLFs is that the features do not require significant design time. However, the increased dimensionality of the feature space leads to many problems, such as: the “curse of dimensionality” [24], increased computational complexity, and possible overfitting due to the sparsity of the feature space. In fields that lack large amounts of data, this sparsity issue can be very problematic as it is hard to show that the classifier is generalizable to new data. Furthermore, classification results using LLFs cannot easily convey intuitive rationale, as the features themselves are not intuitive to a user, leading to reduced clinical acceptance [15].

C. Definition

We defined an HLIF as follows:

A mathematical model that has been carefully designed to describe some human-observable characteristic, and whose outcome can be intuited in a natural (e.g., visual) way.

In contrast to LLFs, HLIFs usually require more upfront design time. An HLIF captures a specific characteristic that is relevant to the given application (e.g., complexity of the color distribution, smoothness of an object), making intuitive feedback possible. As a result, fewer HLIFs may be needed to accurately describe the data. This idea is explored further in Section IV.

D. How to Design an HLIF

The first step in designing an HLIF is to study the target user. The goal is to understand how they analyze the data. Recall that HLIFs are modeled according to a *human-observable characteristic*. These characteristics are unique to each application. This can be accomplished by conducting observational studies or a literature review.

The second step is to identify available tools for modeling high-level characteristics. For example, perceptually uniform color spaces (e.g., CIE $L^*a^*b^*$) can be used to quantify color distribution patterns.

The third step is the modeling stage. The feature should describe a high-level characteristic such that intuitive feedback can be provided to the user (e.g., graphically).

Although this paper focuses on the application of HLIFs for melanoma detection, the HLIF framework can be applied to other problems that involve classification of data using semantic decomposition.

III. SKIN LESION HLIFs

This section presents the design and calculation of ten HLIFs for the detection of melanoma in images obtained using standard consumer-grade cameras. These features were designed to model the intuitive ABCD metric widely used by dermatologists. Since the feature models follow the HLIF framework, the system can provide intuitive diagnostic rationale. The proposed asymmetry features are extensions of the work presented in [18], and the border irregularity features are extensions of the work presented in [19], [20]. The diameter (“D”) criterion was not addressed since the acquisition process was unconstrained, making scale inference challenging.

A. Asymmetry HLIFs

Dermatologists try to identify asymmetry of the shape and/or color of a skin lesion. While benign nevi tend to have homogeneous color distributions, melanomas tend to be asymmetrically pigmented [5], [6]. Furthermore, while benign nevi tend to be elliptically shaped, melanomas tend to have complex shapes.

1) *HLIF for Color Asymmetry*: The goal of an HLIF for describing color asymmetry is to differentiate lesions based on the spatial uniformity and symmetry of the color distribution. This feature is similar to field color asymmetry [25], except that Earth mover’s distance (EMD) is used instead of entropy and many axes of separation are considered.

Given a segmented skin lesion, the major axis was chosen as the initial axis of separation (AoS). The major axis passes through the center of mass (i.e., centroid) of the lesion shape and describes the maximum amount of structural variation (i.e., the transverse diameter of the fitted ellipse). The color distributions in the perceptually uniform CIE $L^*a^*b^*$ space on each side of this AoS were compared. In particular, k “signatures” [26] on both sides of the AoS were determined using k -means clustering, using the final k clusters as color signatures. Mathematically

$$S_i^\theta = k\text{-means}(C_i^\theta, k) \quad (1)$$

where θ denotes the orientation of the AoS, $S_i^\theta \in \{S_1^\theta, S_2^\theta\}$ is the color signature (weighted clusters) in CIE $L^*a^*b^*$ space to either side of the AoS, $C_i^\theta \in \{C_1^\theta, C_2^\theta\}$ is the color distribution to either side of the AoS in CIE $L^*a^*b^*$ space, and $k\text{-means}(C_i^\theta, k)$ is k -means clustering of data C_i^θ into k clusters.

Intuitively, S_i^θ is a set of points in 3-D space, where each point has a mass equivalent to the number of points within the cluster. The EMD [26] was computed using these two signatures. This “distance” metric quantifies the amount of perceptual work needed to transform the color distribution from one side of the lesion to the color of the other, thus effectively representing the amount of color asymmetry. This formulation was repeated over n equally spaced orientations so that a uniform sampling of AoS was considered. The feature calculation was determined to be the maximum asymmetry score yielded over the n trials. We used a fast implementation of the EMD in our calculations [27]. To ensure consistent calculations and enhance user-system trust, we used the deterministic PCA-Part k -means initialization [28], [29].

The final feature calculation is as follows:

$$f_1^A = \max_{\theta} \{ \text{EMD}(S_1^\theta, S_2^\theta) \} \quad (2)$$

where θ is the orientation of the AoS, S_1^θ and S_2^θ are the color signatures in CIE $L^*a^*b^*$ space as in (1). Since dermatologists seek to identify six unique colors using dermatoscopes [5], [6], we used $k = 10$ color clusters to account for the varying lighting conditions present in standard camera images, and $n = 12$ separation axes.

Fig. 1 depicts an example of this HLIF. The maximal AoS is plotted as a white line through the centroid of the lesion, and the obtained CIE $L^*a^*b^*$ color signatures of both sides of the AoS are plotted, where the size of the sphere denotes the number of pixels belonging to that cluster centroid (i.e., weight). Fig. 1(a) and (c) intuitively captures the primary observable colors above and below the AoS. For example, Fig. 1(b) shows primary dominance of light-tan colors as well as smaller concentrations of dark-brown colors. The two color signatures are markedly different, which is captured using the EMD.

2) *HLIF for Structural Asymmetry*: A lesion’s shape becomes less likely to be symmetric as it deviates from the ideal elliptical structure. Structural asymmetry can, therefore, be approximated by the coarse complexity of the lesion’s spatial structure. The lesion’s shape was reconstructed using Fourier descriptors [30] in two coarse manners to quantify structure complexity, according to the following algorithm. This builds on previous features using Fourier descriptors (e.g., [31]

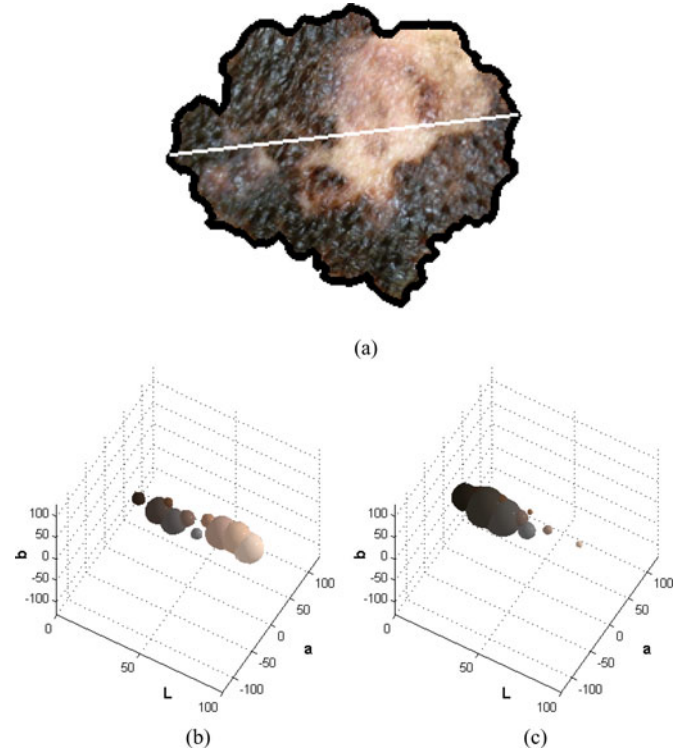


Fig. 1. Example of f_1^A on a superficial spreading melanoma with asymmetric colors. Notice how (b) and (c) capture the intuitive color characteristics of the lesion on each side of the line, irrespective of texture and lighting variation. It is apparent that “work” is required to transform one color signature into the other. In this case, $f_1^A = 23.86$ (a) segmented lesion. (b) $L^*a^*b^*$ color signature above the line. (c) $L^*a^*b^*$ color signature below the line.

and [32]) by adding a reconstruction step to the descriptor process.

The lesion border was sampled using a predetermined sampling rate. This is necessary since the number of frequencies represented by the discrete Fourier transform is directly related to the number of discrete spatial samples. Using a constant sampling rate ensures consistent reconstruction. The Fourier descriptors of the shape were computed. In particular, the fast Fourier transform (FFT) was applied on the complex number $f = x + iy$ where (x, y) are the border pixels’ Cartesian coordinates [30]. In order to omit certain high-frequency information, frequency components were discarded (i.e., their amplitudes were set to 0). The inverse FFT (IFFT) was used to generate two low-frequency reconstructions. The first reconstruction used the lowest two frequencies, which represented the coarsest approximation of the lesion border assuming an elliptical shape. The second reconstruction used $n > 2$ frequencies to capture the presence of coarse structural variability. The normalized area between these two reconstructions was used to quantify the amount of complexity. Complex structures exhibit large area differentials, and simple structures (e.g., elliptically shaped benign lesions) exhibit very little difference.

The final feature calculation is as follows:

$$f_2^A = \frac{\text{area}(R_2 \oplus R_n)}{\text{area}(R_2 \cup R_n)} \quad (3)$$

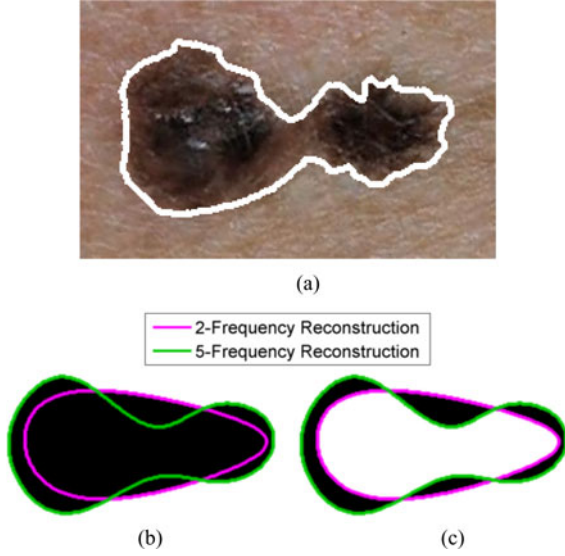


Fig. 2. Example of f_2^A on a superficial spreading melanoma with asymmetric structure. The asymmetry is introduced due to the lack of pigmentation density in the middle of the lesion. This structural variation is captured in the area differential between the two-frequency and five-frequency border reconstructions. In this case, $f_2^A = 0.327$ (a) Segmented lesion (b) Total area (c) Differential area.

where $\text{area}()$ is a function that calculates the area of a geometric shape, R_2 and R_n are the two-frequency and n -frequency reconstructions ($n > 2$), and \oplus, \cup are the XOR and UNION operators, respectively. The XOR area can be interpreted as the differential lesion area as compared to an elliptical lesion. In our tests, we used a 1000-point sampling rate and used five low-frequency components.

Fig. 2 depicts an example of this HLIF using $n = 5$. The structure is very asymmetric, with one side containing much less abnormal pigmentation than the other. The coarse structure variation is captured in Fig. 2 by the five-frequency reconstruction (green) and not the two-frequency reconstruction (pink). Thus, there exists a significant area differential between the two, indicating likely asymmetry. This can be intuitively observed in the differential area plot.

B. Border Irregularity HLIFs

Dermatologists try to identify irregular borders of the skin lesion. Melanoma cases tend to have highly irregular pigmented borders such as “spiky” borders [5], [6].

1) *HLIF for Fine Irregularities:* Melanoma cases often contain abrupt localized pigmentation patterns, such as “spikes.” In order to quantify these “fine” irregularities, the theory of morphological operations [33] can be used. This feature draws from the morphological shape representation theory such as [34].

Morphological operations, unlike Fourier descriptors, are able to manipulate shapes on a *local* scale. The amount of localized abrupt pigmentation can be measured using morphological opening and closing. The resultant normalized difference in area from these operations was compared to the original lesion. This can be measured using the normalized self-dual top-hat operator, described in the following.

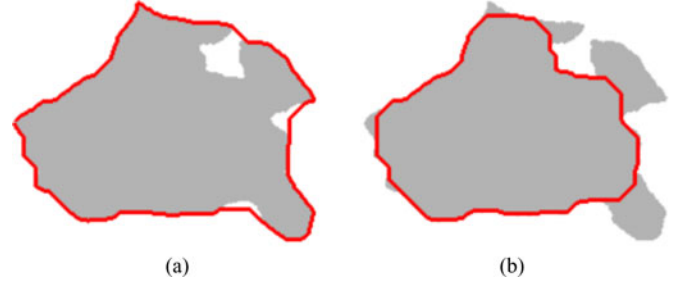


Fig. 3. Example of f_1^B on a lesion shape whose border contains fine irregularities (peaks and valleys). Morphological closing successfully fills in the abrupt valleys, and morphological opening fills out the abrupt peaks. In this case, $f_1^B = 0.208$ (a) Morphological closing (b) Morphological opening.

The final feature calculation is as follows:

$$f_1^B = \frac{T_b + T_w}{A_{\text{lesion}}} \quad (4)$$

where T_b and T_w are the black top-hat and white top-hat morphological operators, and A_{lesion} is the area of the original segmentation. Here, the numerator is the self-dual top-hat operator. Specifically, $T_b = A_{\text{closed}} - A_{\text{lesion}}$, and $T_w = A_{\text{lesion}} - A_{\text{opened}}$, where A_{opened} and A_{closed} are the areas resulting from performing morphological opening and closing on the original lesion. The normalized self-dual top-hat operation represents the amount of exterior and interior irregularities. In our tests, we used a disk structuring element of radius 20.

Fig. 3 depicts an example of this HLIF. In Fig. 3(a), irregular valleys in the border are filled in, accounting for a significant proportion of the overall area. Similarly in Fig. 3(b), the irregular peaks are filled out. The amount of area filled in/out is a good indicator of border irregularities.

2) *HLIF for Coarse Irregularities:* Coarse border irregularities may also be present in melanoma cases. These irregularities are general structural shapes that deviate from an elliptical shape (e.g., notches in the border).

From a signal-processing perspective, these irregularities can be conceptualized as large spatial variations in low-frequency information. Fourier descriptors were used again to capture this information. In particular, much like the structural asymmetry HLIF presented in Section 2), the lesion was sampled at a pre-determined sampling rate and reconstructed using only a small number of low frequencies. To quantify the coarse structural deviations, the perimeters of the low-frequency reconstruction and the original border lesion were compared.

The final feature calculation is as follows:

$$f_2^B = \frac{|P_{\text{orig}} - P_{\text{low}}|}{P_{\text{orig}}} \quad (5)$$

where P_{orig} and P_{low} are the perimeter lengths of the original and low-frequency reconstruction. We used 1000-point sampling rate using linear interpolation and used four-frequency reconstructions.

Fig. 4 depicts an example of this HLIF. Notice how the border contains several points at which it swoops down below and back up above the low-frequency border reconstruction. These are characteristic patterns of coarse border irregularities, where the border does not follow a smooth oval shape.

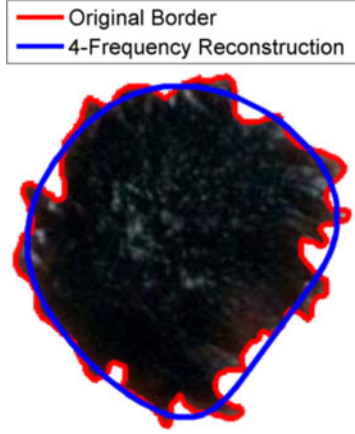


Fig. 4. Example of f_2^B on a superficial spreading melanoma with coarse border irregularities. Notice the general deviations away from the smooth oval shape produced by the four-frequency reconstruction. In this case, $f_2^B = 0.325$.

C. Color HLIFs

Recurring color patterns have emerged in melanoma cases [5]. Unfortunately, most of the ABCD color characteristics [5], [6] are only observable with the aid of a dermatoscope. Furthermore, many image processing tools for medical image analysis were developed for monochrome images [35]. Many existing color features are statistical features in either RGB or alternative color spaces [9]. There is, therefore, a significant demand for novel research on quantifying color information pertaining to melanoma detection, particularly using standard camera images.

The clinical definition of malignant melanomas states that they exhibit varying color patterns, which can be interpreted as complex nonuniform color distributions compared to benign nevi. The spatially varying pigment densities are a result of the metastatic growth of melanocytes. While exact color patterns vary widely, the fundamental “complex” nature of the color distribution can be observed in many melanoma cases.

The goal of these HLIFs is, therefore, to capture the *complexity* of the color distribution. An intuitive way to determine this information is to compare reconstructions of the lesion’s color distribution using fixed numbers of representative colors. The fundamental theory for this framework draws is similar in nature to a recent paper that proposed k -means for color-spatial representation [36]. In addition, it adds a reconstruction stage using this theory for color complexity analysis.

Consider a typical benign nevus with a fairly uniform color distribution. This lesion’s color can be estimated fairly accurately by using one representative color for a given lesion (i.e., a representative red color). In contrast, melanoma cases (see, for example, Fig. 1) exhibit varying colors, making it difficult to find a single color that accurately represents the lesion’s color distribution.

First a color complexity analysis framework is presented, which was used to design the proceeding HLIFs. The color complexity analysis framework is comprised of the following four steps.

Step 1: Transform the image to a perceptually uniform color space.

Step 2: Construct color-spatial representations that model the color information for a patch (i.e., local grid) of pixels.

Step 3: Cluster the patch representations into k color clusters.

Step 4: Quantify the variance found using the original lesion and the k representative colors.

Step 1: Perceptual Uniformity The original RGB image was transformed into the CIE $L^*a^*b^*$ space [37], in which the color distribution is approximately perceptually uniform under the D50 illuminant. Although we cannot assume standard illuminant profile across all images, these effects are reduced if we compare relative perceptual color changes. The perceptual difference should be similar regardless of slight tonal differences between cameras. This way, color values may be compared according to approximate relative perceptual difference.

Step 2: Patch Representation The goal of this step was to represent each patch of pixels in such a way that patches with similar pigmentation get grouped together in Step 3. To do this, two types of information were extracted from each patch: color information and spatial information. This way, spatial constraints enforce locally cohesive color structures, modeling the spatial localization of skin blotches. This can be represented by concatenating each column of pixel values in a patch across each CIE $L^*a^*b^*$ channel into a 1-D vector, and encoding the center pixel coordinate for spatial context. Mathematically, for a given square pixel patch $P_w(\mathbf{x})$ of width w centered around the pixel at location $\mathbf{x} = (\mathbf{x}_x, \mathbf{x}_y)$ in image I in CIE $L^*a^*b^*$ space, the spatial (r^s) and color (r^c) representations for patch P_w were defined as

$$r^s(P_w(\mathbf{x})) = \begin{bmatrix} \mathbf{x}_x & \mathbf{x}_y \end{bmatrix} \quad (6)$$

$$r^c(P_w(\mathbf{x})) = \begin{bmatrix} g_{(L^*)}(P_w(\mathbf{x})) & g_{(a^*)}(P_w(\mathbf{x})) & g_{(b^*)}(P_w(\mathbf{x})) \end{bmatrix} \quad (7)$$

$$g_\lambda(P_w(\mathbf{x})) = \begin{bmatrix} P_w(\mathbf{x}_{11}, \lambda) & P_w(\mathbf{x}_{21}, \lambda) & \dots & P_w(\mathbf{x}_{w1}, \lambda) & P_w(\mathbf{x}_{12}, \lambda) & \dots & P_w(\mathbf{x}_{w2}, \lambda) \end{bmatrix} \quad (8)$$

where $P_w(\mathbf{x}_{ij}, \lambda)$ is the pixel value from channel λ at the i th row and j th column in the patch. Note that $\text{card}(g_\lambda(P_w(\mathbf{x}))) = w^2$, making $\text{card}(r^c(P_w(\mathbf{x}))) = 3w^2$ and $\text{card}(r^s(\mathbf{x}_i)) = 2$, where $\text{card}(\cdot)$ is the cardinality function. A graphical depiction of r^c is given in Fig. 5. The final representation was the concatenation of the spatial and color information

$$r^t(P_w(\mathbf{x})) = \begin{bmatrix} r^c(P_w(\mathbf{x})) & r^s(P_w(\mathbf{x})) \end{bmatrix}. \quad (9)$$

Step 3: Representative Colors Upon populating the projection space with vectors $r^t(P_w(\mathbf{x}_i))$ for each point \mathbf{x}_i in the image, k -means clustering was used to determine the k most representative colors of the lesion. Recall that k -means performs clustering by minimizing the within-cluster sum-of-squares distance of the clusters. This translates to clustering according to perceptual similarity. For consistency and reproducibility, PCA-Part using Otsu’s method was used [28], [29] to generate a deterministic cluster initialization. Since the effect of the spatial characteristics in (6) is affected by the patch size (i.e., the

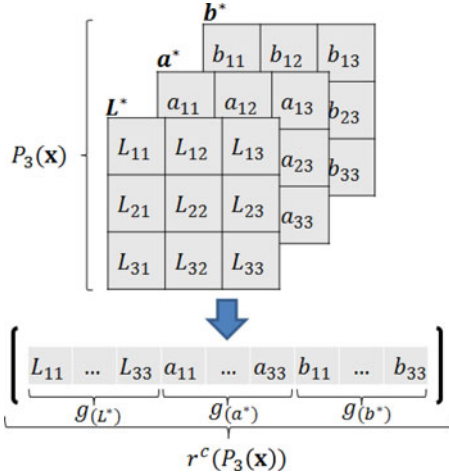


Fig. 5. Graphical representation of the patch color representation in (7). L_{ij} , a_{ij} , b_{ij} represent the pixel values for the particular channel. For each channel, the pixel values are concatenated consecutively by column.

length of the feature vector), an additional weighting term was added into the k -means within-class sum-of-squares criterion as follows:

$$D = \arg \min_S \sum_{j=1}^k \sum_{r^t(\cdot) \in S_j} \left(\|r^c(P_w(\mathbf{x}_i)) - \mathbf{d}_j^c\|^2 + \|\lambda \cdot r^s(P_w(\mathbf{x}_i)) - \mathbf{d}_j^s\|^2 \right) \quad (10)$$

where $D = \{\mathbf{d}_i\}_i$ is a set of k centroid color-spatial elements, S_j is a color cluster, d^c and d^s are the color and spatial components of the color-spatial element, and λ is a relative spatial weighting term. The set $\{\mathbf{d}_i^c\}_i$ can be regarded as a set of representative color patches.

Step 4: Color Reconstruction The output of Step 3 is a set D of k colors (i.e., the centroids of the clusters) along with the pixels in each cluster. Using this information, the image (lesion) can be reconstructed by replacing each pixel's original value with that of the representative centroid. These reconstructions can be used to quantify the amount of color variation.

HLIFs for Color Complexity: Recall that the goal is to use the reconstruction scheme to generate features whose models can be intuited. As aforementioned, "simple" benign lesions may be reconstructed accurately using as little as one representative color, whereas melanoma color distributions are usually more complex. Three sets of HLIFs were generated to satisfy these conditions. In our tests, we used 9×9 patches ($w = 9$), spatial weight $\lambda = 1.75$, and $k = \{1, 2, 5\}$ clusters. These values were chosen such that we could compare against the "base case" (one cluster), as well as more complex reconstructions while noticing that lesion images mostly comprise a few colors but have pixels saturated with lighting artifacts. Dermatologists look for six distinct colors using a dermatoscope [5], [6], however, not all of these colors are observable using standard cameras. We, therefore, chose five colors as our upper bound for reconstruction.

Using this color complexity analysis framework, six HLIFs were constructed that characterize the complexity of a lesion's

color distribution. Each feature calculation can be done following the execution of the color complexity framework.

1) *HLIFs for Quantifying Reconstruction Error:* The first set of HLIFs treats the original lesion as the "ground truth." It compares the relative reconstruction errors in CIE $L^*a^*b^*$ space using one-, two-, and five-color reconstructions. If there is little difference between these reconstructions, it can be concluded that the color distribution is simple; conversely, larger differences indicate more complex color spaces.

The final formulation of these two HLIFs is as follows:

$$f_1^C = \frac{\text{RMSD}(I_{\text{Lab}}, R(I_{\text{Lab}}, D_2))}{\text{RMSD}(I_{\text{Lab}}, R(I_{\text{Lab}}, D_1))} \quad (11)$$

$$f_2^C = \frac{\text{RMSD}(I_{\text{Lab}}, R(I_{\text{Lab}}, D_5))}{\text{RMSD}(I_{\text{Lab}}, R(I_{\text{Lab}}, D_1))} \quad (12)$$

where $\text{RMSD}(I_1, I_2)$ is the root mean squared difference between images I_1 and I_2 , I_{Lab} is the lesion in CIE $L^*a^*b^*$ space, and $R(I, D_k)$ is the color reconstruction of image I using k color patches (using computed set D from Step 3). These HLIFs represent the relative reconstruction error between one-versus-two and one-versus-five color patches, thus capturing the complexity of the color distribution.

Fig. 6 depicts an example of these HLIFs. Reconstruction error using one color (i.e., $\text{RMSD}(I_{\text{Lab}}, R(I_{\text{Lab}}, D_1))$) is large since one cluster is insufficient to reconstruct the complex color distribution of the original lesion. The reconstruction error decreases slightly with the two-color reconstruction, although much of the red and pink pigmentation is still not present. The error is substantially decreased with the five-color reconstruction, which successfully reconstructs the tan, pink, red, and dark pigmentations. The rate of reconstruction error over the number of clusters is quantified using f_1^C and f_2^C .

2) *HLIFs for Quantifying Color Complexity Evolution:* The second set of HLIFs quantifies the amount of color complexity by comparing the evolution of the color distribution across reconstructions with varying numbers of color clusters. Underlying color patterns emerge when reconstructing with more colors, whereas a simple lesion might be accurately represented by only a single color. To capture this information, the mean "difference" between reconstruction using one, two, and five colors was computed. This was computed using the mean ℓ_2 difference between two reconstructions in the CIE $L^*a^*b^*$ space, resulting in a value that portrays the perceptual difference between the reconstructions.

The final formulation of these two HLIFs is as follows:

$$f_3^C = \frac{1}{N} \|R(I_{\text{Lab}}, D_5) - R(I_{\text{Lab}}, D_1)\|_F \quad (13)$$

$$f_4^C = \frac{1}{N} \|R(I_{\text{Lab}}, D_5) - R(I_{\text{Lab}}, D_2)\|_F \quad (14)$$

where N is the number of pixels in the lesion, $\|\cdot\|_F$ is the Frobenius norm, and $R(I, D_k)$ is the reconstruction of image I with k color clusters as in (11) and (12).

Fig. 6 depicts an example of these HLIFs. Treating the one-color reconstruction as the "base case," there is drastic color evolution using two- and five-color reconstructions. For example, the two dominant pigments are tan and black, both of which are reconstructed using two clusters. The less dominant red and

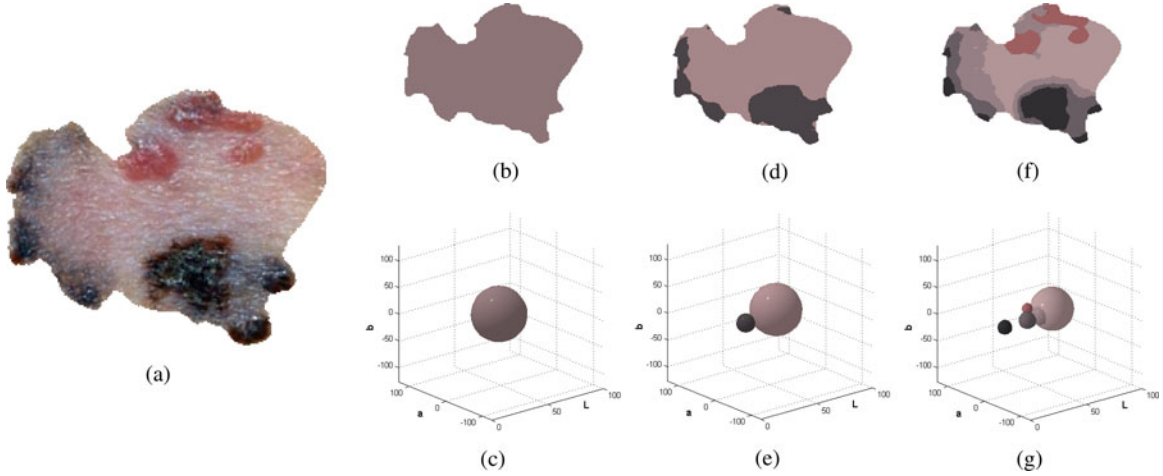


Fig. 6. Example of $\{f_1^C, \dots, f_6^C\}$ on a superficial spreading melanoma with a complex color distribution. Figs. (b), (d), and (f) are color reconstructions using the proposed color complexity analysis framework from Section III-C, and Figs. (c), (e), and (g) are the associated clusters in CIE $L^*a^*b^*$ space. The size of each sphere indicates the number of pixels belonging to that cluster (“mass”). The color complexity is apparent when analyzing the change in reconstructions and color signatures as the number of clusters increases. In this case, $f_1^C = 0.682$, $f_2^C = 0.480$, $f_3^C = 0.189$, $f_4^C = 0.095$, $f_5^C = 125.3$, $f_6^C = 73.9$ (a) Original lesion (b) 1-cluster reconstruction (c) 1-cluster signature (d) 2-cluster reconstruction (e) 2-cluster signature (f) 5-cluster reconstruction (g) 5-cluster signature.

pink pigments are reconstructed using five clusters. This evolution is successfully modeled by the HLIFs.

3) *HLIFs for Quantifying Mean Color Differences*: The third and final set of HLIFs compares the lesion’s color *signatures* across different numbers of clusters. The EMD is again a very appropriate tool for quantifying this information. A “signature” is defined as a cluster representation of the point distribution, where the number of points belonging to each cluster is stored [26]. When comparing two signatures, EMD does not require these signatures to be the same size. This is an important property for our use of color comparison.

EMD calculates how much “work” is needed to transform one signature into another by considering the distance and mass when moving a point from one cluster to another. In our case, the distance is the Euclidean distance in CIE $L^*a^*b^*$ space, and the mass is the number of pixels belonging to a particular cluster after the deterministic k -means procedure.

The final formulation of these two HLIFs is as follows:

$$f_5^C = \text{EMD}(S_1, S_2) \quad (15)$$

$$f_6^C = \text{EMD}(S_2, S_5) \quad (16)$$

where S_k is the signature using k color clusters.

Fig. 6 depicts an example of these HLIFs. The one-cluster signature [see Fig. 6(c)] is a single cluster that approximates the average lesion color. The two-cluster signature [see Fig. 6(e)] incorporated some of the dark pigmentation, effectively creating a smaller black cluster at a perceptual distance away from the original cluster. The five-cluster signature [see Fig. 6(g)] incorporated the red pigmentation, as well as some of the more subtle pink pigments. It can be observed that a nontrivial amount of “work” is required to transform the one-cluster signature to the two-cluster signature by “transporting” the tan pigment to the dark pigment, and that even more work is required to transform the two-cluster signature to the five-cluster signature by creating the red and subtle pink pigments. This is indicative of a lesion with large color complexity.

IV. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the HLIFs proposed in Section III. This feature set was analysed with a state-of-the-art LLF set modeled according to the ABCD rule [12], which is a complete ABCD feature set that was shown to attain higher accuracy than existing full ABCD feature sets [12]. This 52-feature LLF set has been characterised as “low level” according to the definition given in Section II-B. The final proposed feature set was the combined set of HLIFs and LLFs. Analysis was performed in two manners. First, the features were statistically analyzed independently of classification using our dataset to assess the inherent class separability of the features. Second, a standard classification scheme was performed. Finally, observations and limitations of the experimental results are discussed.

A. Data

We collected 206 images of skin lesion, which were obtained using standard consumer-grade cameras in varying and unconstrained environmental conditions. These images were Extracted from the online public databases Dermatology Information System [21] and DermQuest [22]. Of these images, 119 are melanomas, and 87 are not melanoma. Each image contains a single lesion of interest. This is the same dataset used in [18] and [20].

B. Experimental Setup

For each image, the lesion was manually segmented to provide an “ideal” segmentation for feature extraction. That is, we wished to analyze the feature extraction performance irrespective of an automatic segmentation’s accuracy. We rendered the images rotation- and scale-invariant by performing the following preprocessing step: prior to feature extraction, the image was rotated so that the lesion’s major axis was parallel to the horizontal axis, and the lesion fit within a 200×200 bounding

box while maintaining the original aspect ratio. The decision support workflow was implemented in MATLAB.

1) *Preprocessing*: We applied the multistage illumination modeling (MSIM) skin illumination correction algorithm [38]. Briefly, MSIM uses a Markov chain Monte Carlo approach to estimate a nonparametric illumination model of the healthy skin. This model is used as a prior to fit a quadratic surface to the pixels. Finally, this quadratic surface is applied to the computed reflectance map of the image to correct for the lighting variation contributing to the nonuniform skin surface reflection.

2) *Feature Extraction*: Following the illumination correction, the HLIFs presented in Section III were extracted as well as an existing feature set [12], which is the most recent full ABCD feature set designed for standard camera images of pigmented skin lesions, to the best of the authors' knowledge. For simplicity of discussion and analysis, the following notation is used throughout this section.

- 1) S_L —set of 52 LLFs describing ABCD [12].
- 2) S_H —set of 10 HLIFs presented in Section III.
- 3) S_T —set of 62 features obtained by concatenating S_L and S_H (i.e., $S_T = S_L \cup S_H$).

Prior to passing the feature vectors into a classification scheme, feature scaling was performed. Specifically $\forall f \in S_T$, let f_i be feature calculation f for image i . Then, each feature was normalized as [39]

$$f_i^* = \frac{f_i - \mu_f}{\sigma_f} \quad (17)$$

where f_i^* is the normalized feature value, and μ_f and σ_f are the mean and standard deviation over all computed feature scores f_i for feature f . This formulation transforms the data such that each feature exhibits zero-mean and unit standard deviation (and variance) across the dataset. It has been shown that scaling feature vectors eases numerical difficulties in SVM's optimization, and may result in better classification performance [40]. Perhaps of more relevance to this study is that this simplifies the task of determining the feature score's significance, as each feature score distribution is modeled by the statistics of a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. For example, a feature score of $f_i = 2$ signifies that it is two standard deviations away from the mean feature score exhibited by the dataset, representing that it is larger than roughly 98% of the rest of the data. Although this is a simplifying assumption of the data distribution, it serves as an approximate indicator of feature scale.

3) *Classification*: We used a linear soft margin SVM classifier [41] due to its widely regarded robustness and simplicity. The linear kernel was chosen to emphasize the degree of linear separability of the data in the feature space rather than the performance of a complex classifier (which is important in decision support systems, but is out of the scope of the feature extraction stage). Good accuracy can, therefore, be attributed to the feature extraction algorithm's ability to project the data into a separable feature space. We used the LIBSVM implementation for our experiments [42].

There are two parameters that influence the linear soft margin SVM optimization, denoted by c (error cost) and w_i (class i weight) in LIBSVM. To find an accurate SVM hyperplane for

the data, we optimized these parameters in accordance with the LIBSVM authors' recommendations [40] by choosing the best F_β [43] averaged across 50 independent cross-validation trials over a geometric grid search. For each trial, we used a random 80%/20% data split for training/testing. Mathematically

$$F_\beta = \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \quad (18)$$

where

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (19)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (20)$$

where TP, FP, and FN are the number of true positive (correct malignant prediction), false positive (incorrect malignant prediction), and false negative (incorrect benign prediction) cases, respectively. Recall is weighted β -times as important as precision [43]. If the average F_β using (c_i, w_i) was greater than the previous maximum F_β , (c_i, w_i) were stored.

C. Evaluating Classification Accuracy

Due to the lack of large datasets in melanoma detection research, we used the leave-one-out cross-validation (LOO CV) strategy for evaluating the success of the classification. LOO CV is useful when dealing with this problem—that is, evaluating the classifier's ability to generalize using limited data. In particular, for each image's feature vector $\mathbf{f}_k \in S$, the SVM classifier was trained on $S \setminus \mathbf{f}_k$ and tested on \mathbf{f}_k , yielding a binary result: pass or fail. For a dataset with n elements, this strategy resulted in n independent training and testing phases, of which the total error was determined by the total number of incorrect predictions divided by n .

D. Statistical Evaluation of the Feature Space

A statistical analysis of the extracted features was performed to assess class separability. This provides classifier-independent measures of the performance of each feature. Although these tests hold their own unique set of assumptions, conclusions can be drawn on each feature independently, thus contributing to a more complete analysis of the proposed features. The metrics used to evaluate the feature space are described here, along with the computed results.

1) *Two-Sample t -Test*: Given sample data from two classes, a two-sample t -test seeks to reject the null hypothesis H_0 that two class sample distributions come from the same population distribution. It assumes that the classes are normally distributed with unknown but unequal variances according to $\mathcal{N}(\mu, \sigma_i^2)$, where μ is the population mean and σ_i^2 is the class variance. The feature score normalization (17) transforms the feature scores to a zero-mean unit-variance distribution.

Table I shows the results of applying Welch's two-sample t -test for each HLIF. A low p -value indicates that a particular feature differentiates between the two classes (malignant and benign) very well, under the assumption of normality. Three of the features exhibited p -values less than 0.001, indicating that

TABLE I
STATISTICAL SIGNIFICANCE OF EACH HLIF USING TWO TESTS (WELCH'S TWO-SAMPLE t -TEST AND MANN-WHITNEY U TEST)

HLIF	f_1^A	f_2^A	f_1^B	f_2^B	f_1^C	f_2^C	f_3^C	f_4^C	f_5^C	f_6^C
Welch p -value	<0.001	0.036	<0.001	<0.001	0.004	0.039	0.13	0.026	0.048	0.002
U test p -value	<0.001	0.003	<0.001	<0.001	0.003	0.03	0.061	0.007	0.072	0.004

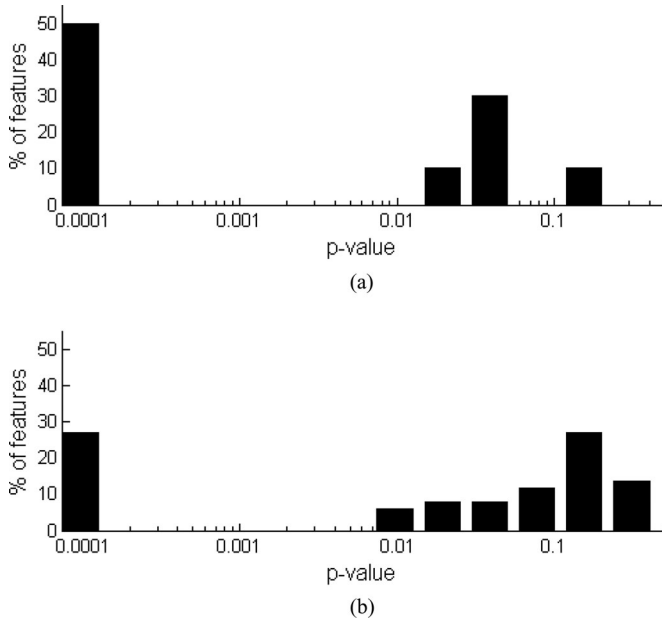


Fig. 7. Comparison of p -values using Welch's two-sample t -test on each HLIF and LLF from [12]. The x -axis is log scale, indicative of p -value importance. A much higher percentage of HLIFs exhibited small p -values than the LLFs, alluding to the discriminating power of HLIFs (a) Histogram of p -value for S_H . (b) Histogram of p -value for S_L .

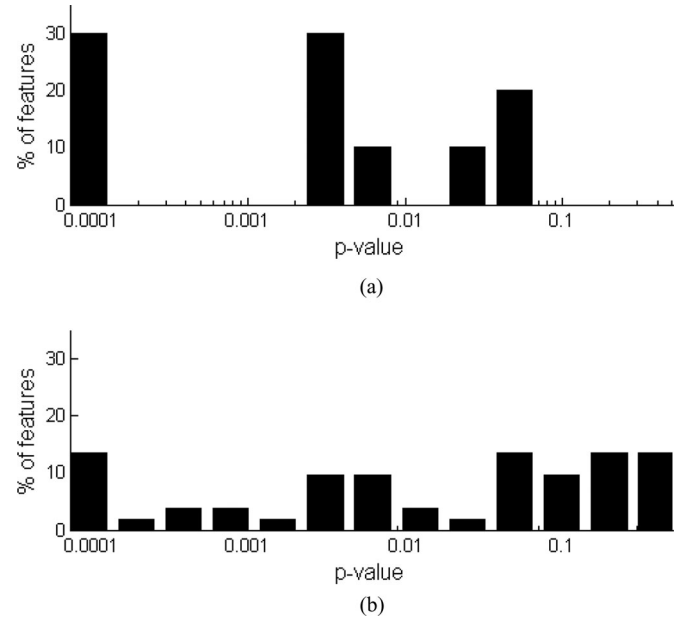


Fig. 8. Comparison of p -values using the nonparametric Mann-Whitney U test on each HLIF and LLF from [12]. The x -axis is log scale, indicative of p -value importance. A higher percentage of HLIFs exhibited small p -values than the LLFs, alluding to the discriminating power of HLIFs (a) Histogram of p -value for S_H . (b) Histogram of p -value for S_L .

they are good feature descriptions. Further, only one feature (f_3^C) exhibited a p -value greater than 0.10.

The normalized histograms of p -values for S_L and S_H using Welch's two-sample t -test are shown in Fig. 7. A larger percentage of features from S_H exhibited low p -values compared to those from S_L . For example, 50% of the features from S_H exhibited p -values less than 0.001, whereas only 28% of the features from S_L exhibited such low p -values. These results allude to the discriminative power of HLIFs for problems that involve improving intuitive identification.

2) *Mann-Whitney U Test*: Another test of data separability is the nonparametric Mann-Whitney U test [44]. Given sample data from two classes, the Mann-Whitney U test seeks to reject the null hypothesis H_0 that two class sample distributions come from the same population distribution. It assumes that the different population have different medians.

Table I shows a similar story as the t -test results: HLIFs do discriminate between the classes for which they are designed. Fig. 8 shows that a much larger percentage of HLIFs attain lower p -values than LLFs. These results are consistent with the results obtained using the two-sample t -test, providing further evidence that HLIFs are very discriminatory especially compared to LLFs.

TABLE II
COMPARING RESULTS OF DIFFERENT FEATURE SETS OVER TEN INDEPENDENT CLASSIFICATION TRIALS

Optimization	Feature set(#)	Sensitivity (%)		Specificity (%)		Accuracy (%)	
		μ^1	σ^1	μ	σ	μ	σ
Function (F_β)							
	F_1						
	$F_{1.25}$						
	S_L (52)	91.43	1.18	66.55	2.39	80.92	0.65
	S_H (10)	92.52	3.75	39.66	9.34	70.19	2.14
	S_T (62)	91.01	1.64	73.45	3.69	83.59	1.14
	$F_{1.25}$						
	S_L (52)	92.94	1.54	64.48	1.38	80.92	0.64
	S_H (10)	96.22	1.78	32.64	2.61	69.37	0.28
	S_T (62)	92.52	1.22	66.09	2.78	81.36	1.15
	$F_{1.5}$						
	S_L (52)	94.37	1.05	57.59	5.24	78.83	1.93
	S_H (10)	96.64	0.56	31.49	2.18	69.13	0.92
	S_T (62)	92.94	1.95	65.06	7.21	81.17	2.03

μ mean, σ standard deviation. See Section IV-B2) for feature set descriptions.

E. Classification Results

The feature sets were evaluated using the classification scheme described in Section IV-B3. Results using the feature sets S_L , S_H , and $S_T = S_L \cup S_H$ are summarized in Table II. For comparison purposes, SVM parameter selection was performed using three different forms of F_β ($\beta = \{1, 1.25, 1.5\}$). The stochastic SVM parameter selection

strategy was randomized using an 80%/20% train/test split for each cross-validation trial. The sensitivity and specificity scores were therefore obtained over ten independent parameter selection and classification runs. The mean and standard deviation of the accuracy metrics across the ten trials were used to show the consistency of the results.

For each F_β , the best mean results in Table II are shown in boldface for each metric (i.e., sensitivity, specificity, and accuracy). The following observations can be made from the results.

1) *General Accuracy Patterns:* S_H consistently attained slightly higher sensitivity metrics than the other two feature sets across each F_β (92.52%, 96.22%, 96.64% for $F_1, F_{1.25}, F_{1.5}$, respectively). However, its specificity was usually low compared to the other two feature sets due to the weight of recall (sensitivity) in the optimization framework. By appending S_L to S_H , S_T consistently attained the highest specificity and accuracy metrics of all the feature sets.

Classification using S_T exhibited higher specificity than S_L in all cases, while attaining high sensitivity with a comparable standard deviation to S_L . This alludes to S_T 's capability of attaining high performance scores with high test–retest reliability. S_H attained high sensitivity with a relatively large standard deviation, indicating moderate test–retest reliability. Appending the LLFs increased the test–retest reliability, however, a larger dataset should significantly abate this shortcoming since the HLIF values would be more indicative of the population distribution. Since HLIFs are designed to model human-observable characteristics, a large dataset is required to “learn” such high-level characteristics. This is discussed further in Section V.

2) *Effect of F_β on Classification Results:* The choice of F_β as an objective function for SVM parameter selection affects the final classification results. Remember that a higher β weighs recall higher than precision during optimization. Since the mathematical formulation of recall is the same as that of sensitivity, as β increases, sensitivity increases and specificity decreases. This is indeed observed in Table II for all three feature sets. This parameter β can be tuned according to the user's preference regarding false positive and false negative rates.

F. Providing Intuitive Rationale

One would expect that a doctor is more likely to trust a computer-generated malignancy prediction if intuitive rationale is provided along with the predicted label. Each HLIF was designed according to the ABCD criteria, which is a visual metric commonly used by dermatologists. To infer intuitive rationale is simple, as each HLIF represents information for which the dermatologist themselves would look. This information can usually be relayed graphically to the user, since melanoma detection is a very visual process.

Fig. 9 shows the 10 HLIF scores for an example image of nodular melanoma. The features had been normalized on the training data so that the significance of a feature calculation could be easily interpreted by the number of standard deviations from the sample mean feature score. Fig. 10 provides an example interface for intuitive visualization of the color asymmetry.

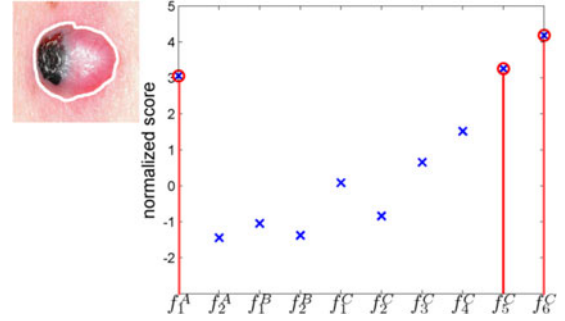


Fig. 9. Example HLIF scores for a nodular melanoma. Abnormally high feature scores are highlighted in red.



Fig. 10. Example intuitive visualization for the case presented in Fig. 9. Upon analyzing the image, the interface indicates that there is apparent color asymmetry and complex color patterns by highlighting the relevant ABCD terms. When the user clicks on “Color,” an overlay is shown to provide intuitive justification for the claim.

Although out of the scope of this paper, this rationale can also be used in a reinforcement learning scheme [45], where the user can give intuitive feedback to evolve the classifier. Furthermore, the output can be used as a learning aid to new or training dermatologists, to test their knowledge of the widely used ABCD rule.

V. DISCUSSION AND LIMITATIONS

One overarching conclusion can be drawn from the experimental results: HLIFs capture relevant information for melanoma detection. We showed that a small set of HLIFs can increase classification performance when combined with a large set of LLFs, and that single HLIFs, if designed properly, are more effective at data separation than single LLFs. The performance of the HLIFs themselves are not fully discriminative, however it was shown in Table I that individual HLIFs capture more discriminative information than LLFs.

A larger dataset may allow the HLIF set to perform better, as the classifier could be trained on more representative class distributions. In its current state, insufficient data leads to suboptimal results.

A large hindrance of the current state of skin cancer detection research is the limited amount of data available to the scientific community [46]. Dermatologists may take pictures of skin lesions, but restrict them to within their clinic, due perhaps to either privacy or commercialization concerns. In order to ensure robust models and statistical validity, much larger datasets must be accumulated for training and testing these decision support systems. This is especially the case since the images are obtained in unconstrained environments, leading to extremely large variations in acquired data.

Another unfortunate by-product of the current data collection methods is that most images presented with a final diagnosis are of late-stage melanoma. Melanoma patients' prognosis is highly correlated with the stage in which it is identified (and excised). Although there is merit in mid- to late-stage melanoma diagnosis, a large emphasis should be placed on early-stage diagnosis for ensuring better survival rates. Again, data collection needs to precede the validation of models for the systems to be accurate.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a framework for designing HLIFs, and has proposed a full set of HLIFs for quantifying skin lesion characteristics for melanoma detection. HLIFs are feature calculations that have been meticulously modeled to describe some human-observable characteristic, and from which rationale can be relayed to the user in some intuitive (perhaps visual) manner. It was shown in Section IV that skin lesion classification accuracy was improved when concatenating a small set of HLIFs and a state-of-the-art LLF set. Individual HLIFs were shown to have more statistical significance with respect to separating the data than individual LLFs in Section IV-D. Furthermore, the addition of HLIFs enabled the provision of intuitive rationale for the predicted malignancy, as shown in Section IV-F.

It is the hope that this study may lead to more clinically viable decision support systems with the aid of HLIF sets. Future work may involve designing HLIFs to characterise melanin and hemoglobin information in RGB images (e.g., [47]), as well as designing HLIFs to model other intuitive criteria, such as EFG for nodular melanoma detection [48], or the 3-D characteristics of the lesion [49]. A much larger endeavor that may have the largest impact on data-deficient melanoma detection field is the systematic construction of a comprehensive dataset.

REFERENCES

- [1] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *Amer. Family Physician*, vol. 62, no. 2, pp. 357–368, 2000.
- [2] R. Lucas, T. McMichael, W. Smith, and B. Armstrong, *Solar Ultraviolet Radiation: Global Burden of Disease from Solar Ultraviolet Radiation* (Environmental Burden of Disease Series), A. Pruss-Ustun, H. Zeeb, C. Mathers, and M. Repacholi, Eds. Geneva, Switzerland: World Health Organization, 2006, no. 13.
- [3] N. Howlader, A. Noone, M. Krapcho, J. Garshell, N. Neyman, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, H. Cho, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin. (2012, Apr.). "SEER cancer statistics review, 1975-2010," National Cancer Institute, Bethesda, MD, USA. [Online]. Available: http://seer.cancer.gov/archive/csr/1975_2010/citation
- [4] American Cancer Society, "Cancer Facts & Figures 2011," American Cancer Society, Atlanta, GA, USA, Tech. Rep. ACSPC-029771, 2011.
- [5] W. Stolz, A. Riemann, A. Cagnetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, M. Landthaler, and O. Braun-Falco, "ABCD rule of dermatoscopy: A new practical method for early recognition of malignant melanoma," *Eur. J. Dermatol.*, vol. 4, no. 7, pp. 521–527, 1994.
- [6] F. Nachbar, W. Stolz, T. Mörkle, A. B. Cagnetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, "The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions," *J. Amer. Academy Dermatol.*, vol. 30, no. 4, pp. 551–559, 1994.
- [7] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives Dermatol.*, vol. 134, no. 12, pp. 1563–1570, 1998.
- [8] Y. Wazaei, A. Tenenhaus, A. Nkengne, J.-F. Horn, A. Giron, S. Paris, and B. Fertil, "The impact of the observation of predictive features on the diagnosis of pigmented skin lesions and the therapeutic decision," presented at the Int. Conf. on Data Mining, Las Vegas, Nevada, USA, 2012.
- [9] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69–90, 2012.
- [10] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 721–733, Sep. 2009.
- [11] H. C. Engasser and E. M. Warshaw, "Dermatoscopy use by US dermatologists: a cross-sectional survey," *J. Amer. Academy Dermatol.*, vol. 63, no. 3, pp. 412–419, 2010.
- [12] P. G. Cavalcanti and J. Scharcanski, "Automated prescreening of pigmented skin lesions using standard cameras," *Comput. Med. Imag. Graph.*, vol. 35, no. 6, pp. 481–491, 2011.
- [13] J. F. Alcon, C. Ciuhu, W. Ten Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, and G. de Haan, "Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 14–25, Feb. 2009.
- [14] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis* (Lecture Notes in Computational Vision and Biomechanics), M. E. Celebi, and G. Schaefer, Eds. New York, NY, USA: Springer, 2012, vol. 6, pp. 63–86.
- [15] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, 2007.
- [16] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, pp. 527–539, 1987.
- [17] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, art. no. 5, Dec. 2002.
- [18] R. Amelard, A. Wong, and D. A. Clausi, "Extracting high-level intuitive features (HLIF) for classifying skin lesions using standard camera images," in *Proc. Conf. Comput. Robot Vis.*, Toronto, Canada, May 2012, pp. 396–403.
- [19] R. Amelard, A. Wong, and D. A. Clausi, "Extracting morphological high-level intuitive features (HLIF) for enhancing skin lesion classification," in *Proc. Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.*, San Diego, CA, USA, Sep. 28, 2012–Sep. 1, 2012, pp. 4458–4461.
- [20] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "Melanoma decision support using lighting-corrected intuitive feature models," in *Computer Vision Techniques for the Diagnosis of Skin Cancer* (Series in BioEngineering), J. Scharcanski and M. E. Celebi, Eds. New York, NY, USA: Springer, 2013.
- [21] Dermatology Information System. (2012). [Online]. Available: <http://www.dermis.net>
- [22] DermQuest.(2012). [Online]. Available: <http://www.dermquest.com>
- [23] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.

- [24] R. E. Bellman, *Dynamic Programming*. New York, NY, USA: Dover Publications, 2003.
- [25] A. Horsch, "Melanoma diagnosis," in *Biomedical Image Processing* (Biological and Medical Physics, Biomedical Engineering), T. M. Deserno, Ed. Berlin, Germany: Springer, 2011, pp. 307–328.
- [26] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.
- [27] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 460–467.
- [28] T. Su and J. G. Dy, "In search of deterministic methods for initializing k-means and gaussian mixture clustering," *J. Intell. Data analysis*, vol. 11, no. 4, pp. 319–338, 2007.
- [29] M. E. Celebi and H. A. Kingravi, "Deterministic initialization of the k-means algorithm using hierarchical clustering," *Int. J. Pattern Recog. Artif. Intell.*, vol. 26, no. 7, pp. 1250018-1–1250018-25, 2012.
- [30] R. Gonzalez and R. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson Education, 2011.
- [31] K. M. Clawson, P. J. Morrow, B. W. Scotney, D. J. McKenna, and O. M. Dolan, "Determination of optimal axes for skin lesion asymmetry quantification," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, 2007, pp. 453–456.
- [32] B. Kusumoputro and A. Ariyanto, "Neural network diagnosis of malignant skin cancers using principal component analysis as a preprocessor," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 1998, vol. 1, pp. 310–315.
- [33] J. Serra, *Image Analysis and Mathematical Morphology*. London, U.K.: Academic, 1982.
- [34] P. Maragos, "Pattern spectrum and multiscale shape representation," vol. 11, no. 7, pp. 701–716, 1989.
- [35] M. E. Celebi and G. Schaefer, Eds., *Color Medical Image Analysis* (Lecture Notes in Computational Vision and Biomechanics). New York, NY, USA: Springer, 2012, vol. 6.
- [36] M. E. Celebi and A. Zornberg, "Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification," *IEEE Syst. J.*, vol. 8, no. 3, pp. 980–984, Sep. 2014.
- [37] "Commission internationale de l'éclairage proceedings 1931," CIE, Vienna, Austria, 1932.
- [38] J. Glaister, R. Amelard, A. Wong, and D. A. Clausi, "MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 1873–1883, Jul. 2013.
- [39] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
- [40] C.-W. Hsu, C.-C. Chang, and C.-J. Lin., (2010). A practical guide to support vector classification. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
- [41] C. W. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, pp. 273–297, 1995.
- [42] C.-C. Chang and C.-J. Lin. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* [Online]. 2(3), pp. 27-1–27-27. Available software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [43] C. Van Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.
- [44] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947.
- [45] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
- [46] G. R. Day and R. H. Barbour, "Automated melanoma diagnosis: Where are we at?" *Skin Res. Technol.*, vol. 6, no. 1, pp. 1–5, 2000.
- [47] A. Madocci, M. S. Drew, M. Sadeghi, and M. S. Atkins, "Intrinsic melanin and hemoglobin colour components for skin lesion malignancy detection," in *Medical Image Computing and Computer-Assisted Intervention (Lecture Notes in Computer Science)*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Heidelberg, Germany: Springer, 2012, vol. 7510, pp. 315–322.
- [48] A. J. Chamberlain, L. Fritschi, and J. W. Kelly, "Nodular melanoma: Patients' perceptions of presenting features and implications for earlier detection," *J. Amer. Academy Dermatol.*, vol. 48, no. 5, pp. 694–701, 2003.
- [49] B. D'Alessandro and A. P. Dhawan, "3-D volume reconstruction of skin lesions for melanin and blood volume estimation and lesion severity analysis," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 2083–2092, Nov. 2012.



Robert Amelard (S'11) received the B.S.E. degree in software engineering in 2011 and the M.A.Sc. degree in systems design engineering in 2013 from the University of Waterloo, Waterloo, Canada, where he is currently working toward the Ph.D. degree with the Vision and Image Processing Lab.

From 2007 to 2010, he held various software engineering internship positions in the fields of geospatial analysis, healthcare, finance, and online advertising, as well as research positions in symbolic computation. His current research interests include medical

image processing and medical device design.

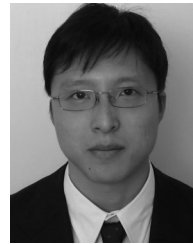
Mr. Amelard received the Alexander Graham Bell Canadian Graduate Scholarship-Doctoral from The Natural Sciences and Engineering Research Council of Canada.



Jeffrey Glaister (S'12) received the B.A.Sc. degree in systems design engineering in 2011 and the M.A.Sc. degree in systems design engineering in 2013, both from the University of Waterloo, Waterloo, Canada. He is currently working toward the Ph.D. degree in electrical and computer engineering at Johns Hopkins University, Baltimore, MD, USA.

He is a member of the Image Analysis and Communications lab, Johns Hopkins University. His current research topic is parcellation of thalamic nuclei from magnetic resonance images and past research

includes segmentation of skin lesions from dermatological photographs. His research interests include biomedical image processing, remote sensing, and pattern recognition.



Alexander Wong (M'05) received the B.A.Sc. degree in computer engineering, the M.A.Sc. degree in electrical and computer engineering, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, Canada, in 2010.

He is currently the Canada Research Chair in Medical Imaging Systems, Codirector of the Vision and Image Processing Research Group, and an Assistant Professor in the Department of Systems Design Engineering at the University of Waterloo. He has published refereed journal and conference papers, as well

as patents, in various fields such as imaging, computer vision, graphics, image processing, multimedia systems, and wireless communications. His research interests include image processing, computer vision, pattern recognition, and cognitive radio networks, with a focus on biomedical and remote sensing image processing and analysis such as image registration, image denoising and reconstruction, image super-resolution, image segmentation, tracking, and image and video coding and transmission.

Dr. Wong received the Outstanding Performance Award, the Engineering Research Excellence Award, the Early Researcher Award from the Ministry of Economic Development and Innovation, two Best Paper Awards by the Canadian Image Processing and Pattern Recognition Society, and the Alumni Gold Medal.



David A. Clausi (S'93–M'96–SM'03) received the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, Canada, in 1996.

Afterward, he worked in Software Medical Imaging at Agfa, Waterloo. He started his academic career in 1997 as an Assistant Professor in geomatics engineering at the University of Calgary, Calgary, Canada. In 1999, he returned to his alma mater and is currently a Professor specializing in the fields of intelligent and environmental systems and was recently the Associate Chair–Graduate Studies. He is an active

interdisciplinary and multidisciplinary researcher. He has an extensive publication record, publishing refereed journal and conference papers in the diverse fields of remote sensing, computer vision, algorithm design, and biomechanics. His research efforts have led to successful commercial implementations including creating, building, and selling his own company.

Dr. Clausi received numerous scholarships, paper awards, and two Teaching Excellence Awards. In 2010, he received the award for Research Excellence and Service to the Research Community by the Canadian Image Processing and Pattern Recognition Society. He was the Cochair of IAPR Technical Committee 7 Remote Sensing during 2004–2006.