EVALUATION OF COLOR SPACES FOR UNSUPERVISED AND DEEP LEARNING SKIN LESION SEGMENTATION

Michael Osadebey, Marius Pedersen and Dag Waaler
Norwegian University of Science and Technology, Norway

ABSTRACT
The reliability of skin cancer diagnosis is dependent on accurate lesion segmentation. The choice of a color space in most contributions on skin lesion segmentation for melanoma detection are based on qualitative rather than quantitative approaches. User experience and theoretical properties of the color space are the two major factors influencing the choice of the color space. For this reason, it may be difficult to optimize segmentation accuracy. This paper evaluates the discrimination power of 5 color spaces and 16 color channels for two unsupervised approaches and a deep learning approach on the segmentation of skin lesion in dermatoscopy images. 600 dermatoscopy images with different levels of cluttering and occluding objects from two different databases were utilized. This study suggests that no single color space or color channel is most suitable in real-world scenarios. Therefore, this study can be regarded as a general framework to select a single or combination of color channels that will enhance the segmentation accuracy of images with different level of scene complexities and illumination variations.

KEYWORDS
Skin Lesion, Color Space, Color Channel, Segmentation, Expectation Maximization, Deep Learning

1. INTRODUCTION
The primary goal of skin lesion segmentation is to identify and separate a lesion region from healthy skin, cluttering and occluding objects. Thus, the ease and difficulty of a segmentation task, as well as the accuracy and efficiency of a segmentation algorithm, is largely dependent on the clarity of the information expressed by color spaces (Lu et al., 2018). Color spaces are digital description of colors for human visual perception of images (Ibraheem et al., 2012).

A researcher's first choice in a segmentation task is a color space which can highlight only the pixels that describe the target object and suppress every pixel belonging to the background. Unfortunately, this ideal expectation may not always be realized. There is yet no color space that has universal application because they are designed with different attributes for different purposes. The discrimination power of a color space is the main reason researchers show preference for specific color space in some segmentation tasks (Yang et al., 2010). The Red, Green and Blue (RGB) color space is a popular color space for acquisition of dermatoscopy images because it is device dependent and particularly suitable for image display (Sanchez et al., 2013). Some contributions, such as (Silveira et al. 2009), prefer the blue channel of the RGB color space because it offers the good discrimination between the lesions and background pixels. However, other contributions such as (Ma et al., 2016) knocked down the use of any RGB channel for image segmentation, because the RGB color space is based solely on chrominance and does not contain separate information on intensity. Under varying illumination condition, it will be difficult to measure the difference between two colors in the RGB color space.

Some contributions advance the HSV (Hue, Saturation and Value) and CIELAB color spaces as having the potential to overcome the shortcomings of the RGB color space. The Hue channel refers to the purity of the color, the Saturation channel quantifies the amount of gray in the color, and Value is the intensity of the color. CIELAB is defined by the International Commission on Illumination (CIE) based on Luminance and the color channels A and B. The luminance channel measures the brightness of the colored image.
Figure 1. A dermatoscopy image in the (a) RGB color space, (b) Value channel of the HSV color space, (c) A channel of the CIELAB color space, (d) Cb component of the YCbCr color space, (e) X component of the CIEXYZ color spaces and (f) the Saturation channel of the HSV color space.

Figure 2. Histogram distribution of six attributes of the 600 test images. (a) Perceptual quality. (b) Spatial extent. (d) Eccentricity. (e) Lesion distance. (f) Level of cluttering and occlusion.

The A and B color channels measures the green-red and the blue-yellow components of the colored image. CIELAB is regarded as a device-independent color space because its colors are independent of how they are created or displayed. The design of HSV and CIELAB color spaces conforms with human perception of colors (Abbas et al., 2013). The HSV and CIELAB color spaces, unlike RGB color space, can separate intensity from color information. This attribute is necessary for pre-processing tasks, such as histogram equalization, required to make correction for illumination variation and shadow attenuation required for skin segmentation. A variant and earlier version of the CIELAB is the CIEXYZ which was introduced in 1931. Figure 1 shows a dermatoscopy image in the RGB color space and color channels extracted from four different color spaces.

Although it is widely acknowledged that accurate and efficient skin lesion segmentation is dependent on the color space (Gomez et al., 2007), there is little or no contribution on how to quantitatively determine the optimal color channel or combination of color channels that will optimize skin lesion segmentation. Current contributions on selection of an optimal color channel can be broadly applied to human skin and face detection (Dang et al., 2017), (Shaik et al., 2015), but these are not application-specific to skin lesion segmentation. This has encouraged heuristics and user experience as the approaches adopted by current techniques on the choice of color space. For this reason, we propose a general framework to quantitatively evaluate the discrimination power of different color spaces and color channels and select the color channel or the combination of color channels that is optimal for a segmentation task that involves a large volume of dermatoscopy images.
2. MATERIALS AND METHODS

2.1 Sources of Data

The experiments for this study utilize 600 dermatoscopy images from two databases. 464 images are from the International Skin Imaging Collaboration (ISIC) archive (https://isic-archive.co). The other 136 images are from the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal (http://www.fc.up.pt/addi). Images from the two databases are single-lesion images acquired with different dimensions. The range of row \( M \) dimensions are \((665 \leq M \leq 6668)\). Corresponding column \( N \) dimensions are \((572 \leq M \leq 4439)\). Each image has ground truth provided by certified dermatologists.

2.2 Dataset Attributes

In order to evaluate the performance of the proposed method on images with different types of artefact, we classify the database images into five categories. The categories and number of images in each category are normal (233), dark corners (84), hair occlusion (93), clutter (57) and multiple artifacts (133). For each image in each category, six visual attributes that influence segmentation accuracy were identified and quantitatively evaluated. Visual attributes are labels that can be given to an image to describe its appearance (Kumar et al., 2011). The first visual attribute is perceptual quality \( D_1 \) which was evaluated using BRISQUE, a no-reference image quality assessment method proposed by (Mittal et al., 2012). Metrics for quantifying the remaining visual attributes are defined in Eq. 1 - 5. Second visual attribute (Eq. 1) is spatial extent \( D_2 \), where \( A_s \) is the total number of pixels that describe the region occupied by the lesion, cluttering and occluding objects in a test image. The metric was normalized using the area of the image. Third attribute (see Eq. 2) is area of lesion \( D_3 \), where \( A_g \) is the number of pixels within lesion region in the ground truth image. Fourth attribute (see Eq. 3) is the lesion position \( D_4 \), where \((x_s,y_s)\) and \((X_g,Y_g)\) are the pixel location of the lesion and image centroid, respectively. Fifth attribute (see Eq. 4) is the eccentricity of lesion \( D_5 \), where \((a, b)\) are the semi-major and semi-minor axes lengths of the lesion. Sixth attribute (see Eq. 5) is the level of cluttering and occlusion \( D_6 \), where \( A_{gs} \) is the total number of pixels that do not belong to either lesion or healthy skin in the test image.

\[
D_2 = 100 \left( \frac{A_s}{MN} \right) \quad (1) \quad D_3 = 100 \left( \frac{A_g}{MN} \right) \quad (2)
\]
\[
D_4 = \sqrt{\left( x_s - X_g \right)^2 + \left( y_s - Y_g \right)^2} \quad (3)
\]
\[
D_5 = \left( 1 - \frac{b^2}{a^2} \right) \quad (4) \quad D_6 = 100 \left( \frac{A_{gs}}{MN} \right) \quad (5)
\]

2.3 Analysis of Dataset Attributes

The test images can be considered as having an acceptable level of perceptual quality. Figure 2a shows that the range of the perceptual quality scores is \((0.47 \leq D_1 \leq 0.96)\). The average perceptual quality score is 0.70. Figure 2b shows that the mean spatial extent is \( D_2 = 23 \) and more than half of the test images had spatial extent that lies in the range\((10 \leq D_2 \leq 30)\). The average area and eccentricity of the lesions, as shown in Figure 2c and Figure 2d, are \( D_3 = 24 \) and \( D_4 = 0.62 \), respectively. The area and eccentricity attributes are indication that a quarter of each image area belongs to the lesion region and the lesion region can be approximated by an ellipse or circle. Figure 2e shows that the mean and range of the lesions distance from the image center are 0.16 and \((0.01 \leq D_5 \leq 0.30)\), respectively. This implies that most of the lesions are located closer to the center of the image, which is common in these kinds of images. The histogram plot in Figure 2f show that about half of the test images can be considered as normal images. The remaining half of the test images contain different levels of cluttering and occluding objects.
2.4 Segmentation Algorithms

We utilize two unsupervised and one deep learning segmentation algorithms in this study. One of the unsupervised techniques is the expectation maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is an iterative procedure to model the parameters that most likely describe the observed data. Each iteration of EM algorithm alternates between the expectation and the maximization steps. The expectation step estimates the missing data given the observed data and current estimate of the model parameters. In the maximization step, the estimate of the missing data in the E-step are used in lieu of the actual missing data followed by maximization of the likelihood function. The likelihood increases at every iteration. The algorithm terminates when there is no significant change between the values at the E-step and the M-step.

The other unsupervised segmentation technique is a three-stage segmentation algorithm proposed by (Guarracino et al., 2019), referred to as SD+. The first step is the conversion of an RGB image into the HSV color space where the saturation channel is extracted. Then detection of confounding factors such as shading and specular reflections. Further, there is an inpainting operation and preliminary segmentation which attenuates the confounding factors and create an initial binary mask. The final step executes morphological post-processing operation, refines the binary mask and extracts the lesion region. Deep learning was implemented using the U-net architecture, a fully convolutional neural network proposed by (Ronneberger et al., 2015) for biomedical image segmentation. A major advantage of the U-net architecture is that it can be trained with relatively few images with reasonable segmentation accuracy. The U-shaped architecture of the network is derived from its contracting and expanding path for implementing encoding and decoding of the input image, respectively. In this study, the encoding was implemented using two 3 x 3 convolutional layers consisting of 64 filters, ReLU and max pooling layer which downsamples the input by a factor of 2. The decoder was implemented using one 4 x 4 transposed convolutional layer consisting of 64 filters which upsamples the input by a factor of 2.

2.5 Implementation

Step 1. Categorization of Images
The data was categorized into 500 training and 100 test data I, where \( t = \{1,2,\ldots,100\} \). Images in both categories were selected to reflect overall attributes of the test data, that is, different types and levels of image artifacts such as dark corners, skin hairs, ruler markings, color chart, gel bubbles and ink marks.

Step 2. Color Space and Color Channel Extraction
Each image, originally acquired in the RGB color space, was transformed to \( K = \{1,2,\ldots,5\} \) color spaces and \( K_c = \{1,2,\ldots,16\} \) color channels. The color spaces are RGB, HSV, CIELAB, YCBCR and XYZ. Three color channels was extracted from each color space, except for the RGB with four color channels, namely red, green, blue and grayscale. The identification codes for each color space and color channel are displayed in Table 1.

Step 3. Lesion Segmentation
The U-net deep learning model was trained on the 500 training images using the stochastic gradient descent with momentum optimizer. The training parameter was set at learning rate \( 1 \times 10^{-3} \), 150 training epochs and batch size of 32. Skin lesions were extracted from each color space and color channel of the 100 test images using the EM and U-net. The SD+ algorithm, because it was designed or optimized for only a single channel, was utilized for only the saturation channel of the HSV color space. For both SD+ and U-net, the number of classes was set to 2 so the algorithms segment the image into the target object and background. Skin lesions, the target object, exhibit unique colors across different color channels. For example, the skin lesion is the darker pixels in the green channel of RGB and it is the brighter pixels in the saturation channel of HSV. This consistency property was exploited to accurately identify and extract the target object from each channel image. Furthermore, the output from EM and U-net were further processed to improve segmentation accuracy.

Step 4. Intra-Image Color Space and Channel Score
For each test image, the similarity between the target object extracted from each color space and each color channel image and the ground truth \( I_g \) is separately measured using the dice similarity index \( d_{t,k} \):

\[
d_{t,k}(I_{t,k}, I_g) = \frac{2|I_{t,k} \cap I_g|}{|I_{t,k}| + |I_g|}
\]
Step 5. Intra-Image Optimal Color Space and Color Channel

For each test image, we collate the dice scores from all the $K$ color space and color channel and form them into a set $\{d_{t,1,2,\ldots,K}\}$ as shown in Eq. (7). We apply the dice score as the parameter to determine the index of the color channel with highest discrimination power for each test image.

$$k = \max_k \{d_{t,1,2,\ldots,K}\}$$  \hspace{1cm} (7)

Table 1. Identification code assigned to each color space and color channel

<table>
<thead>
<tr>
<th>CODE</th>
<th>COLOR SPACE</th>
<th>COLOR CHANNEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RGB</td>
<td>Red</td>
</tr>
<tr>
<td>2</td>
<td>HSV</td>
<td>Green</td>
</tr>
<tr>
<td>3</td>
<td>CIELAB</td>
<td>Blue</td>
</tr>
<tr>
<td>4</td>
<td>YCBER</td>
<td>Grayscale</td>
</tr>
<tr>
<td>5</td>
<td>CIEXYZ</td>
<td>Hue</td>
</tr>
<tr>
<td>6</td>
<td>Saturation</td>
<td>Value</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>Luminance</td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>Value</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>Value</td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td>Value</td>
</tr>
<tr>
<td>11</td>
<td>Y</td>
<td>Value</td>
</tr>
<tr>
<td>12</td>
<td>CI</td>
<td>Value</td>
</tr>
<tr>
<td>13</td>
<td>CR</td>
<td>Value</td>
</tr>
<tr>
<td>14</td>
<td>X</td>
<td>Value</td>
</tr>
<tr>
<td>15</td>
<td>Y</td>
<td>Value</td>
</tr>
<tr>
<td>16</td>
<td>Z</td>
<td>Value</td>
</tr>
</tbody>
</table>

Step 6. Inter-Image Mean Color Space and Channel Score

The dice score for each color space and color channel, computed in step 4, from all the test images are collated. The collated data are used to compute the mean dice score $\bar{d}_k$ for each color space and color channel.

$$\bar{d}_k = \frac{1}{T} \sum_{t=1}^{T} d_{t,k}$$  \hspace{1cm} (8)

The mean dice scores for each color space and each color channel:

$$\bar{d} = \{\bar{d}_1, \bar{d}_2, \cdots, \bar{d}_K\}$$  \hspace{1cm} (9)

The mean of all the mean dice scores $\bar{d}$ of each of all the color space and each color channels:

$$\bar{d} = \frac{1}{K} \sum_{k=1}^{K} \bar{d}_k$$  \hspace{1cm} (10)

Step 7. Histogram of Optimal Color Space and Color Channel

The index $k$ of the optimal color space and color channel, computed in step 5, are formed into an histogram. Let $f_k$ be the number of times a color channel $k$ becomes the optimal color channel across all the test images. The probability $p_k$ that a color channel is the optimal color channel across the test images:

$$p_k = \frac{f_k}{T}$$  \hspace{1cm} (11)

where $\sum p_k = 1$

Step 8. Select the Optimal Color Space and Color Channel

We determine the optimal color channel image from the weighted sum of $p_k$, the probability that a color channel is the optimal color channel across the test images and $d_k$, the mean dice score with minimum distance relative to the mean dice score $\bar{d}$ of all the color channels:

$$I_k = \max_k \left\{ \lambda_1 p_k + \lambda_2 (1 - \|d_k - \bar{d}\|) \right\}$$  \hspace{1cm} (12)

where $\lambda_1$ and $\lambda_2$ are arbitrary weights.

3. RESULTS

Histogram distribution of dice score recorded by the SD+ algorithm is displayed in Figure 3. Figure 4a and Figure 4c shows, using the EM algorithm, the histogram distribution of the number of times a color space and color channel was determined to be the optimal color channel of a test image, respectively. For the same EM algorithm, the mean dice score for each color space (blue color) and the mean dice score (purple color) of all
the dice scores are shown in Figure 4b and Figure 4d, respectively. For deep learning based on the U-net architecture, the histogram distribution of the number of times a color space and color channel was determined to be the optimal color channel of a test image are displayed in Figure 5a and Figure 5c, respectively. The mean dice score of each color space (blue color) and the mean dice score (purple color) of all the dice scores for the deep learning algorithm are shown in Figure 5b and Figure 5d, respectively.

4. DISCUSSION

The SD+ algorithm operates only in the saturation channel of the HSV color space. It recorded a dice score lower than 0.80 for 10 percent of the images. For about 30 percent of the images there was a dice score less than 0.9. This result suggests that SD+ is optimized using the saturation channel of the HSV color space. Although it can be considered as robust, there is room for improvement. The algorithm does not demonstrate the same level of robustness across all the test images. If other operating characteristics have not been optimized, there may be need to consider the combination of other channels to enhance segmentation accuracy.

Figure 3. Histogram distribution of dice scores for unsupervised segmentation using SD+ (Guarracino et al., 2019)

Although the EM algorithm is powerful, its performance can be limited by presence of clutters and occluding objects. Unlike the SD+ algorithm, the EM algorithm does not incorporate a hair removal pre-processing step. For this reason, it will perform poorly in the presence of cluttering and occluding objects (Majtner et al., 2016), (Le et al., 2016). This is evident in Figure 4b where the average dice score for all the color spaces is \( P < 0.6 \). There is remarkable improvement in the performance of the EM algorithm when its operation changes from color space to other color channels (Figure 4c and Figure 4d). Color channels that demonstrated remarkable performance are the red, green, blue, grayscale, saturation, luminance and Y channels. As shown in Figure 4c, the blue channel can be considered the optimal color channel because it has a greater number of times \( P \approx 0.3 \) where it was determined as the optimal color channel of a test image. The fact that no color channel has \( P = 0.5 \) in Figure 4c suggests that no single color space can be considered as most suitable where images with different level of scene complexities and illumination variations are acquired for specific task.

Figure 5b shows that the deep learning algorithm demonstrated robust performance for all the color spaces except for the CIELAB and YCbCr color spaces. The HSV color spaces is the optimal color space because, as shown in Figure 5a, it has more number of times \( P \approx 0.4 \) than other color spaces, where it was deemed the optimal color channel. The deep learning algorithm did not repeat the same level of performance when it was operated using color channels, except for the Z color channel. Comparison of the results in Figure 5b and Figure 5d suggest that the deep learning algorithm performs better with more information provided by color spaces.

5. CONCLUSION

Skin lesion segmentation is an important image analysis task because the accuracy of subsequent steps leading to the diagnosis of skin cancer is dependent on accurate segmentation. This study proposes a general framework to evaluate how color spaces and color channels impact the segmentation of skin lesions.
Figure 4. Expectation maximization algorithm (Dempster et al., 1977). (a) Histogram distribution of the number of times a color space was determined to be the optimal color channel of a test image. (b) The mean dice score of each color space (blue color) and the mean dice score (purple color) of all the maximum dice scores. (c) Histogram distribution of the number of times a color channel was determined to be the optimal color channel of a test image. (d) The mean dice score of each color channel (blue color) and the mean dice score (purple color) of all the maximum dice scores in (c).

Figure 5. Deep learning segmentation using U-net architecture (Ronneberger et al., 2015): (a) Histogram distribution of the number of times a color space was determined to be the optimal color channel of a test image. (b) The mean dice score of each color space (blue color) and the mean dice score (purple color) of all the maximum dice scores. (c) Histogram distribution of the number of times a color channel was determined to be the optimal color channel of a test image. (d) The mean dice score of each color channel (blue color) and the mean dice score (purple color) of all the maximum dice scores in (c).
The study suggests that (1) No single color space can be considered as most suitable where images with different level of scene complexities and illumination variations are acquired for specific task. (2) The discrimination power of a color channel is dependent on the level of clutters and occluding objects that are present alongside the target object. The outcome of this study can be applied in the design of a framework for automatic selection of optimal color channels to enhance the performance of segmentation algorithms on tasks that involves images with different level of scene complexities and illumination variations. Limitations of the study is that the training and testing data may not be reasonably large to optimize the segmentation and that the training parameters of the U-net may not result in optimal segmentation. Further, the simplified post-processing applied to the output of the EM and U-net may not optimize segmentation. Nevertheless, since same parameter was applied throughout the experiments, these limitations does not diminish the general concepts and potential applications of this study.

ACKNOWLEDGEMENT

M. Osadebey is supported by the European Research Consortium on Informatics and Mathematics (ERCIM).

REFERENCES


