IMPROVING DERMOSCOPY IMAGE ANALYSIS USING COLOR CONSTANCY

Catarina Barata, Jorge S. Marques

ISR, Instituto Superior Técnico
Lisboa, Portugal

M. Emre Celebi

Louisiana State University
Shreveport, USA

ABSTRACT

Several methods have been proposed to detect melanomas in dermoscopy images. However, most of these methods are tuned to specific acquisition conditions. That is, their performance is affected when the acquisition setup changes or when data comes from multiple sources. This is what happens with EDRA database that contains images acquired in three different hospitals. In this paper, we discuss the use of color compensation techniques that try to reduce the influence of the acquisition setup on the color features extracted from the images. We show that color compensation provides a significant improvement on the performance of two different systems.

***Index Terms*—** Dermoscopy, Skin Cancer, Color Constancy, Color Features, Computer Aided Diagnosis, Color Detection

1. INTRODUCTION

Several Computer Aided Diagnosis (CAD) systems have been proposed to analyze dermoscopy images and diagnose skin cancer [1, 2, 3, 4]. Although several groups reported highly accurate results, a comparison is difficult because each group uses its own dataset, obtained using specific acquisition devices and illumination conditions. It is well known that changes in the acquisition setup can alter the colors of an image, as can be seen in the examples of Fig.1. The human brain is able to compensate for this variability, but the same cannot be said of a CAD system since such variability introduces changes in the values of color-related features like color histograms or mean color vectors. These features are used in most CAD systems and their dependence on the acquisition setup makes the system less robust and more prone to errors if it has to analyze images that come from multiple sources.

This problem has been identified by several research groups, who have investigated methods to normalize the colors of the acquired images [5, 6, 7, 8, 9]. However, most of these methods either require knowledge about the acquisition system (*e.g.*, the type of camera) [5, 6, 7, 8] or a training (calibration) step [9]. We are interested in exploring a simpler direction that does not require knowledge of the acquisition system properties nor a training step and can deal with data from multiple and unknown sources. Color normalization is a very common problem in computer vision and image processing. A popular strategy to normalize image colors is by the use of Color Constancy methods that dispense a calibration step. We consider an algorithm based on low-level image features: Shades of Gray [10]. A recent survey [11] shows that, despite being very simple, the performance of this method is comparable to the performance of more complex algorithms based on training. Furthermore, it is fast and easy to implement.

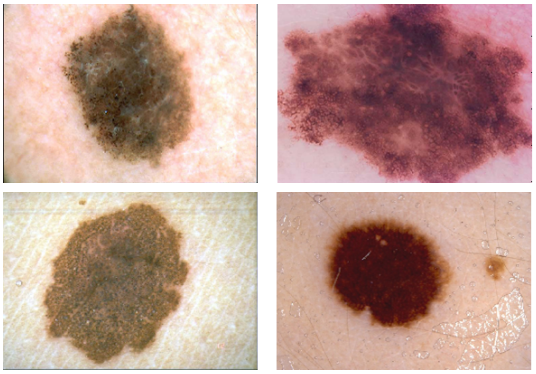

Fig. 1. Examples of images acquired under different illumination conditions: melanomas (1st row) and benign lesions (2nd row) [12].

In this paper we show that the aforementioned method can be used to improve the performance of different dermoscopy CAD systems. First, we show that it significantly improves the performance of a Bag-of-Features (BoF) classifier [13, 14], when applied to a dataset of images from different sources. Then, we show the importance of including color constancy in a CAD system recently proposed to detect relevant colors in dermoscopy images [15]. To the best of our knowledge, this compensation method has not been applied to dermoscopy images and the influence of color calibration in the performance of CAD systems has not been assessed. The remaining paper is organized as follows. Section 2 describes the tested color normalization method. Sections 3 and 4 describe the setups used to evaluate the performance of the

C. Barata and J. S. Marques were partially funded with grant SFRH/BD/84658/2012 and by the FCT project PEst-OE/EEI/LA0009/2013. M. E. Celebi was supported by a grant from the U.S. National Science Foundation (1117457).

color constancy method and present the experimental results. Finally, Section 5 concludes the paper.

2. COLOR CONSTANCY

The goal of color constancy methods is to transform the colors of an image I, acquired under an unknown light source, so that they appear identical to colors under a canonical light source. This task is accomplished by performing two separate steps: estimation of the color of the light source in RGB coordinates, $\mathbf{e} = [e_R\ e_G\ e_B]^T$, and transformation of the image using the estimated illuminant.

Different algorithms have been proposed to estimate the color of the light source [11]. In this work we apply the Shades of Gray method [10], since it does not required prior knowledge about the acquisition setup, and only uses image information to estimate the color of the light source. Furthermore, this method has been shown to achieve similar performances to more complex algorithms that require a training step [11].

For a color image I, each component of the illuminant $e_c,\ c \in \{R, G, B\}$ is estimated using the Minkowski norm as follows

$$\left(\frac{\int (I_c(\mathbf{x}))^p \mathrm{d}\mathbf{x}}{\int \mathrm{d}\mathbf{x}} \right)^{1/p} = k e_c \ \ c \in \{R, G, B\} \ \ , \qquad (1)$$

where I_c is the c-th color component of image I, $\mathbf{x} = (x, y)$ is the position of each pixel, k is a normalization constant that ensures that $\mathbf{e} = [e_R, e_G, e_B]^T$ has unit length, using the Euclidean norm, and p is the degree of the norm. The value of p can be tuned according to the dataset used, as well as according to the type of system. Shades of Gray is a generalization of two other color constancy algorithms: Gray World [16], when $p = 1$, and max-RGB [17], when $p = \infty$.

After estimating **e**, the next step is to transform the image I. A simple way to model this transformation is the von Kris diagonal model [18]

$$\begin{pmatrix} I_R^t \\ I_G^t \\ I_B^t \end{pmatrix} = \begin{pmatrix} d_R & 0 & 0 \\ 0 & d_G & 0 \\ 0 & 0 & d_B \end{pmatrix} \begin{pmatrix} I_R^u \\ I_G^u \\ I_B^u \end{pmatrix} \ \ , \qquad (2)$$

where $[I_R^u, I_G^u, I_B^u]^T$ is the pixel value acquired under an unknown light source, and $[I_R^t, I_G^t, I_B^t]^T$ is the pixel value transformation as it would appear under the canonical light source, which is assumed to be the perfect white light, i.e., $\mathbf{e}^w = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$. The matrix coefficients $\{d_R, d_G, d_B\}$ are the mapping parameters, which are related to the estimated illuminant **e** as follows

$$d_c = \frac{1}{e_c}, \ \ c \in \{R, G, B\} \ \ . \qquad (3)$$

Color constancy is performed on images that are in the sRGB format. Therefore, we perform a gamma correction step ($\gamma = 2.2$) [19] before estimating the illuminant and correcting the image.

We evaluate the influence of color normalization in two different dermoscopy applications. The first one is a CAD system that diagnoses skin lesions using the BoF model [20] and color histograms as discriminative features. On the second case, we assess the importance of color constancy on the development of a CAD system that detects and counts the number of colors in dermoscopy images, using Gaussian mixtures. Both applications are heavily dependent on color features.

3. LESION CLASSIFICATION

BoF has already been applied with success to the classification of dermoscopy images [13, 14] (see the references for further details). The images used in the previous works are the ones of the publicly available PH^2 dataset [21], which were acquired at a single hospital and, thus, using the same equipment and illumination conditions. Our goal is to extend the BoF CAD system to multiple acquisition sources, as can happen in the case of teledermoscopy, where the images are acquired at multiple facilities and sent to a central dermatology service to be diagnosed.

3.1. Experimental Setup

The procedure used to train and test the algorithm is similar to the one described in [13, 14]. In the training step, we start by dividing the training lesions in small patches using the Harris Laplace keypoint detector and characterize them using color histograms. Then, we perform a clustering step to find the centroids (visual words) of all training features. Finally, we assign each patch to the closest visual word, count their occurrence and build a histogram of visual words for each image. To learn the lesions' labels (melanoma/benign) we train a Support Vector Machine (SVM) with the χ^2 kernel

$$K_{\chi^2}(\mathbf{x}, \mathbf{y}) = e^{-\rho d_{\chi^2}(\mathbf{x}, \mathbf{y})} \ \ , \qquad (4)$$

where **x** and **y** are image histograms, ρ is a width parameter, and

$$d_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i} \ \ . \qquad (5)$$

In the test phase, we build a histogram for each new image using the previously computed set of visuals words and classify it using the trained SVM.

In this work we use 1-D RGB histograms to describe the patches. Previous studies have shown that these features lead to good classification results [13, 14]. Furthermore, color constancy algorithms are performed on the RGB color space. In the sequel, we evaluate the BoF classifier with and without color constancy.

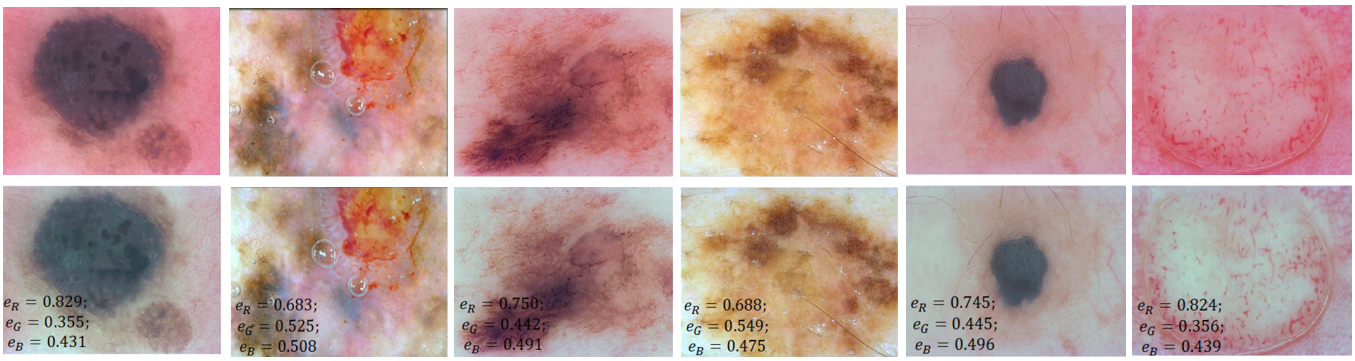

Fig. 2. Color constancy examples - Melanoma (1st-3rd columns) and Benign (4th-6th columns). Original Image (top row), and Shades of Gray $p = 6$ (bottom row).

3.2. Experimental Results

The heterogeneous dataset used in this work was randomly selected from the EDRA database [12], which contains images from three university hospitals: University Federico II of Naples (Italy), University of Graz (Austria) and University of Florence (Italy). We have selected a total of 482 images, 241 melanomas and 241 benign lesions.

To assess the influence of the color calibration methods we have trained two BoF systems. The first was trained using non-normalized images and the other using the color corrected images. To evaluate the performance of each system we compute three metrics: Sensitivity (SE), Specificity (SP) and Accuracy (ACC). SE corresponds to the percentage of melanomas that are correctly classified and SP is the percentage of correctly classified benign lesions. All metrics are computed using a stratified 10-fold cross validation scheme.

Table 1 shows the classification results achieved with and without color corrections. By inspecting this table one can noticed that, as expected, the color constancy method significantly improves the classification results, improving the classification accuracy by 14%. These results were obtained with $p = 6$ in (1).

Table 1. Classification results for the EDRA (multiple source) dataset without and with color constancy correction.

COLOR CONSTANCY	Sensitivity	Specificity	Balanced Accuracy
None	71.0%	55.2%	63.1%
Shades of Gray	79.7%	76.0%	77.8%

Fig. 2 shows some examples of normalized images These examples clearly show that Shades of Gray alters the images, making them look more similar and correcting color channel saturations, like the ones exemplified in the 1st, 3rd, 5th and 6th columns. In these examples, the red channel of the acquisition camera is clearly saturated, making the image look reddish. Color normalization also allows a better visualization of the colors of the lesion, like in the example of the 5th columns where the peripheral light brown area becomes more noticeable, enhances the contrast inside the lesion (see 2nd column) and in some cases even improves the contrast between the lesion and the surrounding skin (*e.g* 3rd, 4th and 6th columns).

Table 2. Classification results for the PH^2 (single source) dataset without and with color constancy correction.

COLOR CONSTANCY	Sensitivity	Specificity	Balanced Accuracy
None	92.5%	75.6%	84.1%
Shades of Gray	92.5%	76.3%	84.3%

We repeated these tests using the PH^2 database with 200 images (40 melanomas and 160 non-melanomas). This database was acquired in single hospital with the same acquisition setup and we wish to determine what is the effect of color constancy corrections on the classification of data from a single source. The results are shown in Table 2. It is concluded that color constancy corrections do not degrade the performance of the classifier when data comes from a single source (there is even a marginal improvement of specificity).

4. COLOR DETECTION

Color plays a major role in the diagnosis of skin lesions. Dermatologists not only screen the lesion for multiple colors, which is a sign of melanoma, but they also consider that some colors are more malignant than others (e.g. white and blue-gray) [12]. Thus, medically oriented CAD systems should include an algorithm to identify and/or quantify the colors in skin lesions.

Recently, Barata et al. [15] proposed an approach to segment and quantify the colors in skin lesions using Gaussian mixtures. Their framework consists of separately learning five Gaussian mixture models to describe five medically relevant colors (light and dark brown, black, blue-gray, and white). Each model is given by

$$p(\mathbf{y}|c,\theta^c) = \sum_{m=1}^{k_c} \alpha_m^c \; p(\mathbf{y}|c,\theta_m^c) \;\; , \qquad (6)$$

where $\mathbf{y}$ is the observed color vector, c denotes the color label, k_c is the number of components of the $c-th$ color mixture,

$\alpha_1^c ... \alpha_{k_c}^c$ are the mixing probabilities ($\alpha_m^c \geq 0 \sum_{m=1}^{k_c} \alpha_m^c = 1$) and $\theta_m^c = (\mu_m^c, R_m^c, \alpha_m^c)$ is the set of parameters that defines the m-th component of the $c-th$ Gaussian mixture. Each mixture is trained using the mean color of regions that have been segmented by an expert. Then, the learned models are applied to a new set of images. Each image is divided into small patches of 12×12 using a regular grid. For each patch its mean color $\mathbf{y}$ is computed and its membership to each color model is determined using a Bayesian approach

$$p(c|\mathbf{y}) = \frac{p(\mathbf{y}|c, \widehat{\theta}^c)p(c)}{p(\mathbf{y}|\widehat{\theta})} \quad , \tag{7}$$

where $\widehat{\theta}^c = (\widehat{\mu}^c, \widehat{R}^c, \widehat{\alpha}^c)$, $\widehat{\theta} = (\widehat{\theta}^1, ..., \widehat{\theta}^5)$, $p(c) = 1/5$ is set to be equal for all colors, and

$$p(\mathbf{y}|\widehat{\theta}) = \sum_{c=1}^{5} p(\mathbf{y}|c, \widehat{\theta}^c)p(c) \quad . \tag{8}$$

By selecting the model with the highest degree of membership it is possible to segment the different color regions. Finally, the number of colors is obtained by comparing the area of each region with a threshold.

Our goal with this second experiment is to determine if the color detection system benefits from this color correction step. In [15] the authors use the HSV color space to compute the color models. We will also use this color space in our color detection experiments. This strategy allows us to understand if color constancy can improve the performance of other color spaces besides RGB, as proposed in [22].

4.1. Experimental Results

To assess the influence of color constancy in the color detection problem we have trained two different sets of color models, using 29 images from the PH^2 dataset (the same ones used in [15]), with medical segmentations of colors. In the first case we train the models without performing the color normalization step, while in the second case we train the models with a color normalization step, where we correct the RGB images before converting them to HSV. We have empirically determined that $p = 3$ (see (1)) is a suitable value.

To test the color detection systems we use two sets: one of 123 images from the PH^2 dataset (different from the training examples) and other with 340 images from EDRA. Please note the mismatch between training and test sources in the second case. For each of these images there is a medical ground truth annotation (color label) stating whether each color is present or absent. The computed statistics are the Sensitivity (SE), percentage of correctly identified colors, Specificity (SP), percentage of correctly non-identified colors, and Accuracy (ACC).

Tables 3 and 4 show the performances obtained with the two tested frameworks. These results show that the system performance is significantly improved when the color constancy algorithm is applied. This conclusion is valid for both datasets. We stress that in the case of Table 4, the models were trained with 29 images from the PH^2 data set and tested with images from a different dataset (EDRA). This shows the robustness of the proposed method.

Table 3. Average results for the detection of the five colors in the PH^2 dataset without and with color constancy correction.

COLOR CONSTANCY	Sensitivity	Specificity	Balanced Accuracy
None	79.7%	73.9%	76.8%
Shades of Gray	91.7%	74.5%	81.7%

Table 4. Average results for the detection of five colors in the EDRA dataset without and with color constancy correction.

COLOR CONSTANCY	Sensitivity	Specificity	Balanced Accuracy
None	75.5%	55.8%	65.7%
Shades of Gray	80.5%	69.7%	75.1%

5. CONCLUSIONS

A color based CAD system for the analysis of dermoscopy images must be able to cope with multi-source images, acquired using different setups and illumination conditions. In this paper we have studied the applicability of color constancy, using two different color-based approaches for dermoscopy image analysis. To apply color constancy we have selected the Shades of Gray method. Nonetheless, we have tried other color constancy methods (Gray World [16] and max-RGB [17], which are special instances of Shades of Gray) and they achieved similar results.

In the lesion classification model, we have shown that color constancy improves the performance of a BoF model in the classification of multi-source images by 14%. When the data is produced by a single source, no improvement or degradation was observed when we use color constancy methods.

The color detection problem is another challenging task. We have investigated the performance of a recently proposed color detection algorithm [15] with and without color normalization and observed significant improvements even when the data is produced by a single source. Color correction seems to enhance color discrimination.

In this work we have represented color using RGB and HSV color spaces, and showed that color constancy significantly improves the performance of the tested systems in both cases. Our results for color detection have shown that, despite color constancy is applied to the RGB images, it can also be used to improve the performance of other color spaces, like HSV.

6. ACKNOWLEDGMENTS

We would like to thank to Prof. Teresa Mendonca and Pedro Ferreira from Faculdade de Ciências, Universidade do Porto, and to Dr. Jorge Rozeira and Dr. Joana Rocha from Hospital Pedro Hispano for providing the annotated PH^2 dataset.

7. REFERENCES

[1] M E. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, Y. Aslandogan, W. Stoecker, and R. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, pp. 362–373, 2007.

[2] H. Iyatomi, H. Oka, M E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, and K. Ogawa, "An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," *Computerized Medical Imaging and Graphics*, vol. 32, no. 7, pp. 566–579, 2008.

[3] J S. Marques, C. Barata, and T. Mendonça, "On the role of texture and color in the classification of dermoscopy images," in *IEEE 34th EMBC*, 2012, pp. 4402–4405.

[4] Q. Abbas, M. E. Celebi, I. F. Garcia, and W. Ahmad, "Melanoma recognition framework based on expert definition of abcd for dermoscopic images," *Skin Research and Technology*, vol. 19, no. 1, pp. e93–e102, 2013.

[5] Y V. Haeghen, J M A D. Naeyaert, and I. Lemahieu, "An imaging system with calibrated color image acquisition for use in dermatology," *IEEE Transactions on Medical Imaging*, vol. 19, no. 7, pp. 722–730, 2000.

[6] C. Grana, G. Pellacani, and S. Seidanari, "Pratical color calibration for dermoscopy applied to a digital epiluminescence microscope," *Skin Research and Technology*, vol. 11, pp. 242–247, 2005.

[7] P. Wighton, T K. Lee, H. Lui, D. McLean, and M S. Atkins, "Chromatic aberration correction: an enhancement to the calibration of low-cost digital dermoscopes," *Skin Research and Technology*, vol. 17, pp. 339–347, 2011.

[8] J. Quintana, R. Garcia, and L. Neumann, "A novel method for color correction in epiluminescence microscopy," *Computerized Medical Imaging and Graphics*, vol. 35, pp. 646–652, 2011.

[9] H. Iyatomi, M E. Celebi, G. Schaefer, and M. Tanaka, "Automated color calibration method for dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 35, pp. 89–98, 2011.

[10] G. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *In Proceedings of IS&T/SID 12th Color Imaging Conference*, 2004, pp. 37–41.

[11] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational color constancy: Survey and experiments," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2475–2489, 2011.

[12] G. Argenziano, H P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, V. Hofmann-Wellenhog, D. Massi, G. Mazzocchetti, M. Scalvenzi, and I H. Wolf, "Interactive atlas of dermoscopy," 2000.

[13] C. Barata, M. Ruela, T. Mendonça, and J. S Marques, "A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*, J. Scharcanski and M. E. Celebi, Eds., pp. 49–69. Springer, 2014.

[14] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *accepted for publication in IEEE Systems Journal*, 2013.

[15] C. Barata, M.A.T. Figueiredo, M.E. Celebi, and J. S. Marques, "Color identification in dermoscopy images using gaussian mixture models," in *Accepted for presentation at ICASSP2014 - BioImaging and Signal Processing*, 2014.

[16] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin institute*, vol. 210, pp. 1–26, 1980.

[17] E. Land, "The retinex theory of color vision," *Scientific America*, vol. 237, pp. 108–128, 1977.

[18] J. von Kries, "Influence of adaptation on the effects produced by luminous stimuli," *Sources of Color Vision*, pp. 109–119, 1970.

[19] Charles Poynton, *Digital video and HD: Algorithms and Interfaces*, Morgan Kaufman, 2012.

[20] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[21] T. Mendonca, P. M Ferreira, J. S Marques, A RS Marcal, and J. Rozeira, "Ph2 - a dermoscopic image database for research and benchmarking," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[22] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *In ECCV*, 2006, pp. 246–268.