Computers in Biology and Medicine 65 (2015) 54–68

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Exploring the color feature power for psoriasis risk stratification and classification: A data mining paradigm

Vimal K. Shrivastava [a], Narendra D. Londhe [a,c], Rajendra S. Sonawane [b], Jasjit S. Suri [c,d,*]

[a] Electrical Engineering Department, National Institute of Technology, Raipur, India
[b] Psoriasis Clinic and Research Centre, Psoriatreat, Pune, Maharashtra, India
[c] Skin Point-of-Care Division, Global Biomedical Technologies, Inc., Roseville, CA, USA
[d] Electrical Engineering Department, Idaho State University (Aff.), Pocatello, ID, USA

ARTICLE INFO

Article history:
Received 10 June 2015
Accepted 29 July 2015

Keywords:
Psoriasis
Color space
PCA
Classification
Feature power
Reliability

ABSTRACT

A large percentage of dermatologist's decision in psoriasis disease assessment is based on color. The current computer-aided diagnosis systems for psoriasis risk stratification and classification lack the vigor of color paradigm. The paper presents an automated psoriasis computer-aided diagnosis (pCAD) system for classification of psoriasis skin images into psoriatic lesion and healthy skin, which solves the two major challenges: (i) fulfills the color feature requirements and (ii) selects the powerful dominant color features while retaining high classification accuracy.

Fourteen color spaces are discovered for psoriasis disease analysis leading to 86 color features. The pCAD system is implemented in a support vector-based machine learning framework where the offline image data set is used for computing machine learning offline color machine learning parameters. These are then used for transformation of the online color features to predict the class labels for healthy vs. diseased cases. The above paradigm uses principal component analysis for color feature selection of dominant features, keeping the original color feature unaltered. Using the cross-validation protocol, the above machine learning protocol is compared against the standalone grayscale features with 60 features and against the combined grayscale and color feature set of 146.

Using a fixed data size of 540 images with equal number of healthy and diseased, 10 fold cross-validation protocol, and SVM of polynomial kernel of type two, pCAD system shows an accuracy of 99.94% with sensitivity and specificity of 99.93% and 99.96%. Using a varying data size protocol, the mean classification accuracies for color, grayscale, and combined scenarios are: 92.85%, 93.83% and 93.99%, respectively. The reliability of the system in these three scenarios are: 94.42%, 97.39% and 96.00%, respectively. We conclude that pCAD system using color space alone is compatible to grayscale space or combined color and grayscale spaces. We validated our pCAD system against facial color databases and the results are consistent in accuracy and reliability.

1. Introduction

Psoriasis is a chronic and autoimmune disease with red and scaly plaques on skin [1]. Statistics show that psoriasis affects about 125 million people of the world population [2]. The influence of psoriasis differs depending on the geographical regions. The prevalence of psoriasis in Europe, USA, Malaysia and India is about 0.6–6.5% [3], 3.15% [3], 3% [4] and 1.02% [5], respectively. Psoriasis not only affects the skin, but also the quality of life [6]. This is due to physical appearance and evil eyes of other people. This results in an increased risk of suicidal attempts (about 30%) [7], making psoriasis an equally dangerous disease and at par with depression, heart disease and diabetes [8]. Thus, psoriasis has become a major concern to health care organizations and the society. Presently, there is no permanent cure for this disease but it can be controlled by prolonged and attentive treatment. The cause of psoriasis is still unidentified but most researchers agree that the prime reason for this disease is genetics [9]. There are various types of psoriasis namely plaque, guttate, inverse, pustular, and erythrodermic. Among all, plaque is most commonly found in about 80% of the cases [10]. Thus, this study is focused on plaque type psoriasis.

Dermatologists generally follow visual inspection and sense of touch to predict the severity. Skilled training and experience is required to infer the disease stage and grade and then to prescribe medication to control this disease. This process is very subjective and brings unreliability to the process of decision making. The process is

* Corresponding author at: Skin Point-of-Care Division, Global Biomedical Technologies, Inc., Roseville, CA, USA. Tel.: +1 916 749 5628.
E-mail addresses: lky.vml@gmail.com (V.K. Shrivastava), nlondhe.ele@nitrr.ac.in (N.D. Londhe), drrajss@gmail.com (R.S. Sonawane), jsuri@comcast.net (J.S. Suri).

http://dx.doi.org/10.1016/j.compbiomed.2015.07.021
0010-4825/

laborious and very time consuming and creates large volume of data sets with huge images hard to handle manually. Thus, the process is very inefficient. They prefer to look for color of the lesions, scales deposited on them and their area of coverage to understand the severity of the diseases. The color of lesion varies from light red to dark red (purplish). The light red indicates reduced scales and improved condition of the disease while dark red indicates large deposition of scales and severe condition. So, hypothetically, color of lesion is adapted as a criterion for classification of psoriasis disease. Our hypothesis is that there is a strong power in color information in psoriasis skin images. When psoriasis images are acquired, they are color images and there is no way to qualitatively tell and classify it because it is all visual. Thus, there is a clear need for an objective analysis combined with subjective analysis. A general review on psoriasis and its risk stratification using CAD system can be seen in our recent review [11].

Many researchers have proposed computer-aided diagnosis (CADx) systems for classification of the dermatology disease [12–14]. These CADx systems either use only grayscale features or combination of color and grayscale features for dermatological disease classification but never tried to understand the power of color features embedded in these colored dermatology images. This paper brings to the forefront the key aspects *i.e.*, color features and psoriasis disease classification as their basis. But, before color features can be selected, the principal color space has to be identified. There are many different color spaces that have been defined in the literature, which may be beneficial in different applications. This paper brings a new paradigm of color spaces with extensive search from literature and then adapted in data mining paradigm for psoriasis disease classification to mimic the dermatologist. So, we explored 14 different color spaces which may be useful for psoriasis disease classification. Further, we investigate the power of color features against grayscale features independently and in a combined framework.

This study brings significant contribution in the field of dermatology by exploring 14 kinds of color spaces for risk assessment using machine learning paradigm. These color spaces namely are: RGB, normalized-RGB, YCbCr, HSV, HSI, CIE XYZ, CIE Lab, CIE Lch, CIE Luv, hunter-Lab, SCT, opponent, CMY and CMYK. Two statistics, *i.e.*, mean and standard deviation is computed for each color space component. Overall, 86 color features are extracted using 14 color spaces. Furthermore, 60 grayscale features are extracted from eight different texture feature extraction techniques namely, Gray-Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Intensity Histogram (IH), Invariant Moment (IM), Gray Level Difference Statistics (GLDS), Neighborhood Gray Tone Difference Matrix (NGTDM), Statistical Feature Matrix (SFM) and Fractal Dimension (FD).

We present an in-depth analysis of the power of color features and grayscale features independently for psoriasis disease classification in support vector machine framework. The reader will find the following new components in this paper: (i) design and exploration of 14 color spaces (86 features) applied standalone for psoriasis color skin disease classification; (ii) design and analysis of 60 grayscale features (derived using 8 grayscale feature extraction techniques) applied indendently for psoriasis color skin disease classification; (iii) ability to choose dominant features using PCA-based feature selection without altering their respective feature values; (iv) introducing the color feature power term which helps in understanding the power of the color spaces in machine learning paradigm, and finally; (v) introducing a reliability factor for CADx systems in enviornments when color and grayscale spaces are independent and combined.

Classification is performed for five different kernel functions *i.e.*, linear kernel, radial basis function (RBF) kernel and polynomial kernel of order one, two and three, where kernel function of polynomial of order two showed the best results. The 10-fold cross-validation protocol is used in the support vector machine-based framework. The proposed system using color features alone shows the highest classification accuracy of 99.94% with sensitivity and specificity of 99.93% and 99.96% using a 10-fold cross-validation using polynomial kernel of order two. Using the color features alone, grayscale features alone and combined color and grayscale features, the sytem accuracy was: 92.85%, 93.83% and 93.99%, respectively. We measure the reliability of the system in these three frameworks yielding 94.42%, 97.39% and 96.00%, respectively. Our results show that the color features alone is as powerful as grayscale features for classificaion of psoriasis disease. Our pCAD results are consistent both in accuracy and reliability while changing the database from psoriasis to facial skin.

2. Psoriasis data acquisition and preparation

The preparation of database is done by mimiking dermatologists view of manually segregating the healthy skin (normal) and psoriatic lesion (abnormal skin) from the images of each patient. The normal and abnormal skin samples are cropped freehand to capture any shape through MATLAB. The images are zoomed to maximum extend to make sure that the border visibility is very clear. As the features calculated depends on the color and texture, so utmost care has been taken to crop lesions from available images. The cropping is done inner to the actual boundary to avoid false inclusion of background which may alter the features generated from the lesion. Thus 270 samples of normal skin and 270 samples of abnormal skin are acquired from the images of 30 patients. There are few samples which are fuzzy in nature *i.e.*, considered as an abnormal sample but actually they belong to normal sample and vice-versa. First three rows of Figs. 1 and 2 show the samples of normal and abnormal skin respectively and last row of Figs. 1 and 2 show the fuzzy normal and abnormal samples respectively. This database has been adapted previously on our earliear developments [15].

3. Methodology

The data mining paradigm is adapted to explore the color feature power. The spirit of this machine learning framework and risk stratification comes from methods developed by Suri and his team [16–19]. This utilizes the machine learning paradigm in support vector machine framework. The machine learning is adapted to establish the generalization using training-based color image data set and a priori physician classified labels, which in turn is used for transforming the test color image data set to predict the unknown risk class. Even though, this model is used for understanding the color space power independently, it can still be extended for grayscale feature set independently, or in a combined way where best color space features and grayscale features can be used. Our extensive amalgamation of 86 color space features and 60 grayscale features utilizes the PCA-based feature selection criteria where the feature set values are unaltered ensuring the selected features are dominant while retaining the stability and reliability of the pCAD system. Our pCAD system uses machine learning paradigm shown in Fig. 3. It has two components as shown by the dotted line. The left side reflects the offline system since it uses the training image data while the right side reflects the online system since it uses the testing image data set. In the offline system, the input RGB image is converted into 13 different color spaces and then features are extracted from different color spaces including RGB color space. In the feature selection step, dominant features were selected using Principal Component