Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2017 12 (1): 133-138

Automatic Classification of Skin Cancer Using KNN, SVM and CNN

Hema Rajini Narayanan

Received: 31 December 2016 ▪ Revised: 10 January 2017 ▪ Accepted: 24 January 2017

Abstract: A skin cancer classification system has been designed and developed. This work presents a new approach to the automated classification of skin cancer images based on texture and colour features. To remove the unwanted noises in the skin image, median filtering is used. In the next stage gray level co-occurrence matrix and colour features are extracted. Finally, k-nearest neighbour, support vector machine and convolutional neural networks are used to classify the skin cancer images. The application of the proposed method for tracking skin cancer is demonstrated to help pathologists distinguish its type of skin cancer. A classification with an accuracy of 85%, 96% and 98% has been obtained by, k-nearest neighbour support vector machine and convolutional neural networks.

Keywords: Support Vector Machine, k-nearest Neighbour, Convolutional Neural Networks.

INTRODUCTION

The largest organ of the body is skin. Skin controls the temperature of the body and stores fat, water and vitamin D. Skin protects us against the sunlight, heat, injury and infection. Skin which contain several layers, but the two main layers are the epidermis (upper or first layer) and the dermis (lower or inner layer). Skin consists of three different types of cells. There is squamous cell, basal cell, melanocytes. The different cells are present on the epidermis layer. The first layer of the epidermis made up of squamous cells, the cell is a thin and flat cell. The central part of the squamous cell is basal cell; the cell is a round cell. The lower part of the epidermis layer is melanocytes cells, that make melanin. Melanin is the pigment that gives skin its natural colour. When skin is exposed to the sun, melanocytes make more pigment. It causes skin to darken.

Skin cancer is rapidly increasing in the world. Survival rate of skin cancer is high, if it is detected early stage. So, an efficient method is necessary to detect skin lesion at the earliest. The cost of dermatoscope scan for screening the patient is high. The main aim of a skin cancer detection system is to reduce the percentage of error by choosing the appropriate method in each stage.

There are many computer-aided classification systems for skin cancer images in the literature, most of them are used to detect and classify abnormalities. Jeya Ramya et al., proposed an active contour and Support Vector Machine (SVM) to detect the melanoma skin cancer [1]. The main aim is to reduce the percentage error using different method in each stage. For pre-processing wiener filter and adaptive histogram equalization is used to enhance the quality of the image and remove unwanted illumination of the image. Segmentation is carried out by active contour model to segment the skin lesion from the pre-processed image. After segmentation the features are extracted by using gray level co-occurrence matrix and classify the skin lesion using SVM. The performance metric such as sensitivity, specificity and accuracy are 90%, 85% and 95%.

Mengistu et al., has proposed a Self Organizing Map (SOM) and Radial Basis Function (RBF) to diagnosis the skin cancer [2]. The input image is pre-processed by using median filter to remove unwanted noise, bubbles, enhance the image and fine hair in the image. The enhanced image is extracted by using GLCM, morphological and colour features. The extracted features are given as the input to the SOM to classify the skin cancer. The results are compared with different classifiers such as KNN, ANN, naïve classifier and SOM. Among them SOM is efficient. The overall accuracy is 93.15%.

Gohila Vani et al., used a method to segment and classify the stages of skin cancer [3]. To remove noise the input image is pre-processed by using dilation and edge detection algorithm. After pre-processing the image is segmented by using Texture Distinctiveness Lesion Segmentation (TDLS) algorithm and features

Hema Rajini Narayanan, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamilnadu, India. E-mail: auhemasmith@yahoo.co.in

are extracted by GLCM. The extracted features are classified by using probabilistic neural network to classify the skin lesion.

Dua et al., developed a method to distinguish basal cell carcinoma (BCC) from benign skin lesions, although the patterns that separate the two are nonobvious [4]. Artificial neural networks (ANNs) may be good pattern classifiers for this application. The result shows the potential of neural networks to distinguish benign from malignant skin lesions using electrical impedance is presented. Electrical impedance was measured in vivo from 1 kHz to 1 MHz at five virtual depths on 18 BCC and 16 benign or premalignant lesions. A feed-forward neural network was trained using back propagation to classify these lesions. Neural networks were able to classify measurements in a test set with 100% accuracy for the first pre-processing technique and 85% accuracy for the second.

Elgamal used a KNN and ANN method to classify the objects as normal or abnormal skin cancer [5]. The input image is pre-processed by median filter to remove noise, uneven illumination and extract the features vectors using Discrete Wavelet Transform (DWT). After obtaining the feature vectors the Principal Component Analysis (PCA) is used to reduce the data dimension. Finally, these feature vectors are classified using ANN and KNN classifier to classify the skin cancer.

Ballerini et al., proposed non-melanoma skin lesion classification using colour image data in a hierarchical k-NN classifier [6]. The accuracy of the proposed hierarchical scheme is higher than 93% in discriminating cancer and pre-malignant lesions from benign lesions, and it reaches an overall classification accuracy of 74% over five common classes of skin lesion.

The rest of this paper is organized as follows. Section 1 presents the introduction as well as the studies of several research papers are portrayed. Section 2 presents the proposed technique, utilized in this work for classification of skin cancer images. In this section, pre-processing, feature extraction and classification are presented. Section 3 experimentally demonstrates the performance of the proposed method. Finally, Section 4 describes the conclusion of this paper.

MATERIALS AND METHOD

Materials

In order to perform automatic classification, a set of skin cancer image was required, and the class label where it belongs to, 225 dermoscopic skin cancer images were taken from DERMOFIT from the predefined three types of skin cancer i.e. melanoma, basal cell carcinoma and squamous cell carcinoma.

Proposed Method

This proposed research work is used to improve skin cancer classification in skin images. The proposed method has three stages, namely pre-processing, feature extraction and classification. The proposed technique for automatic skin cancer image classification is illustrated in Figure 1. The proposed system is developed using Matlab (The Math Works, Inc., Natick, MA, USA). In the first stage, noise is suppressed using an image filtering. In the second stage gray level co-occurrence matrix and colour-based features are extracted. Finally, KNN, SVM and CNN classifiers are used to classify the type of lung tumor images.

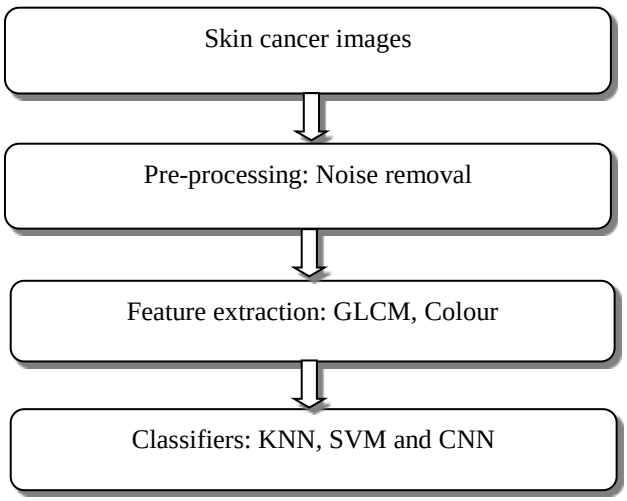

Fig. 1: Methodology of the Proposed Technique

a) Pre-processing

To reduce noise, median filtering using a 3-by-3 square kernel is applied [7]. Median filter is chosen because it is less sensitive to extreme values and able to remove outliers without reducing sharpness of the image. This produces a more homogeneous background in which abnormalities become more conspicuous.

b) Texture Features from Gray Level Co-occurrence Matrix

Texture is a repeating pattern of local variations in image intensity. The co-occurrence matrix is a statistical method used for texture analysis. As the name suggests, the co-occurrence matrix is constructed from the image by estimating the pair wise statistics of pixel intensity. The use of the co-occurrence matrix is based on the hypotheses that the same gray-level configuration is repeated in a texture. This pattern will vary more by fine textures than by coarse textures. The co-occurrence matrix *P(i, j|d, θ)* counts the co-occurrence of pixels with gray values *i* and *j* at a given distance *d* and in a given direction *θ.* According to the number of intensity points (pixels) in each combination, statistics are classified into first-order, second-order and higher-order statistics. In the first order, texture measures are statistics calculated from an individual pixel and do not consider pixel neighbor relationships. The gray level co-occurrence matrix (GLCM) method is a way of extracting second order statistical texture features [8]. However, the performance of a given GLCM based feature, as well as the ranking of the texture features; depend on the number of gray levels used. The following notations are μ is the mean value of P. μ_x, μ_y, σ_x and σ_y are the means and standard deviations of P_x and P_y. G is the size of the co-occurrence matrix. Here the number of rows and columns of the co-occurrence matrix is equal. The following GLCM based texture features are extracted in this research work: contrast, correlation, energy, homogeneity and entropy. They are defined in eqs. (1)-(5).

Contrast

$$Contrast = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G} \sum_{j=1}^{G} P(i,j) \right\}, |i-j| = n \quad (1)$$

Contrast is a measure of the local variations present in an image. This measure of contrast favors contributions from *P (i, j)* away from the diagonal, i.e. *i= j*. If there is a large amount of variations in an image, the *P[i,j]*'s will be concentrated away from the main diagonal and the contrast will have a high value.

Correlation

$$Correlation = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \quad (2)$$

Correlation is a measure of gray level linear dependence between the pixels at the specified positions relative to each other. The correlation will be higher if an image contains a considerable amount of linear structure.

Energy

$$Energy = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j)^2 \quad (3)$$

The energy of a texture describes the uniformity of the texture. Energy is 1 for a constant image.

Homogeneity

$$Homogeneity = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i,j)}{1+|i-j|} \quad (4)$$

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Homogeneity is 1 for a diagonal GLCM. A homogeneous image will result in a co-occurrence matrix with a combination of high and low *P [i,j]*'s. A heterogeneous image will result in an even spread of *P [i,j]*'s.

Entropy

$$Entropy = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j) \times log\big(P(i,j)\big) \quad (5)$$

Entropy statistic measures the disorder or complexity of an image. Complex textures tend to have high entropy. Entropy is strongly, but inversely correlated to energy.

Colour Features

Colour is one of the features of skin cancer, they have different colour variation of each cancer type and colour analysis is computed by taking the mean value of RGB (Red, Green and Blue) components and the mean value of HSIs (Hue, Saturation and Intensity) components. Therefore, to compute the mean value of each component of these colour spaces, matlab is used to split each component because matlab has a built-in function to convert to HIS or RGB colour spaces. By using a function RGB colour image stack is split to red, green and blue components. Hence, the colour features are extracted by computing the mean values of RGBs and HSIs of Dermoscopy skin cancer images. That is, the mean value of red, mean value of green, mean value of blue, mean value of hue, mean value of saturation and mean value of intensity are computed from each component.

c) Classifiers: KNN, SVM and CNN

Classification is the process of classifying the given input by training with a suitable classifier. Deep learning and Support Vector Machine (SVM) classifiers are the best classifiers suggested by many researchers which can be opted for the skin cancer classification of dermoscopic images. It is independent of dimensionality and feature space. Convolutional Neural Networks (CNN) is one of the most remarkable approaches of deep learning, in which multiple layers of neurons are formed in a robust manner. In this work simple classifier like KNN is also used.

k-Nearest neighbor classifier: One of the simplest classification techniques is the k-nearest neighbor (k-NN) classifier. Classification of the input feature vector X is done by determining the k closest training vectors according to a suitable distance metric. Vector X is then assigned to that class to which the majority of that k-nearest neighbors belong. The k-NN algorithm is based on a distance function and a voting function in k-nearest neighbors; the metric employed is the Euclidean distance measure [9]. The k-NN classifier is a conventional nonparametric supervised classifier that is said to yield good performance for optimal values of k. Like most learning algorithms, k-NN algorithm consists of a training phase and a testing phase. Data points are given in an n-dimensional space in the training phase. The labels associated with the data points designate their class in the training phase. In the testing phase, unlabeled data are given and the algorithm generates the list of the k-nearest (already classified) data points to the unlabeled point. This classifier returns the class of the majority of that list.

Support vector machine (SVM) is a powerful supervised classifier and accurate learning technique. From the statistical theory it was derived and developed by Vapnick in 1982. It yields successful classification results in various application domains, e.g. medical diagnosis. SVM is based on the structural risk minimization principle from the statistical learning theory [10]. The kernel controls the empirical risk and classification capacity in order to maximize the margin between the classes and minimize the true costs. SVM searches an optimal separating hyper-plane between members and non-members of a given class in a higher dimensional feature space. The inputs to the SVM algorithm are the features extracted using the GLCM and colour features. In this method, three classes are used.

Convolutional neural networks: CNNs achieve better classification accuracy on large scale datasets due to their capability of joint feature and classifier learning [11]. The convolutional layer plays a vital role in the operation of CNN. The layers parameters focus around the use of learnable kernels. These kernels are usually small in spatial dimensionality, but spreads along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map.

The fully-connected layer contains neurons of which are directly connected to the neurons in the two adjacent layers, without being connected to any layers within them. This is analogous to way that neurons are arranged in traditional forms of ANN. Pooling layers reduce the dimensionality of the representation, and thus further reduce the number of parameters and the computational complexity of the model.

RESULTS AND DISCUSSION

To validating the results of the proposed method, a benchmark image database is employed which comprises 225 skin cancer images. It was taken from DERMOFIT from the predefined three types of skin cancer i.e. melanoma, basal cell carcinoma and squamous cell carcinoma.

In the first stage, noise is suppressed using an image filtering. In the second stage, five texture features using gray level co-occurrence matrix and colour features are extracted. In this research work, five texture and colour features are extracted. They are contrast, correlation, energy, homogeneity and entropy. Finally, KNN, SVM and CNN classifiers are used to classify the type of skin cancer images. Figure 2 shows the input skin cancer images.

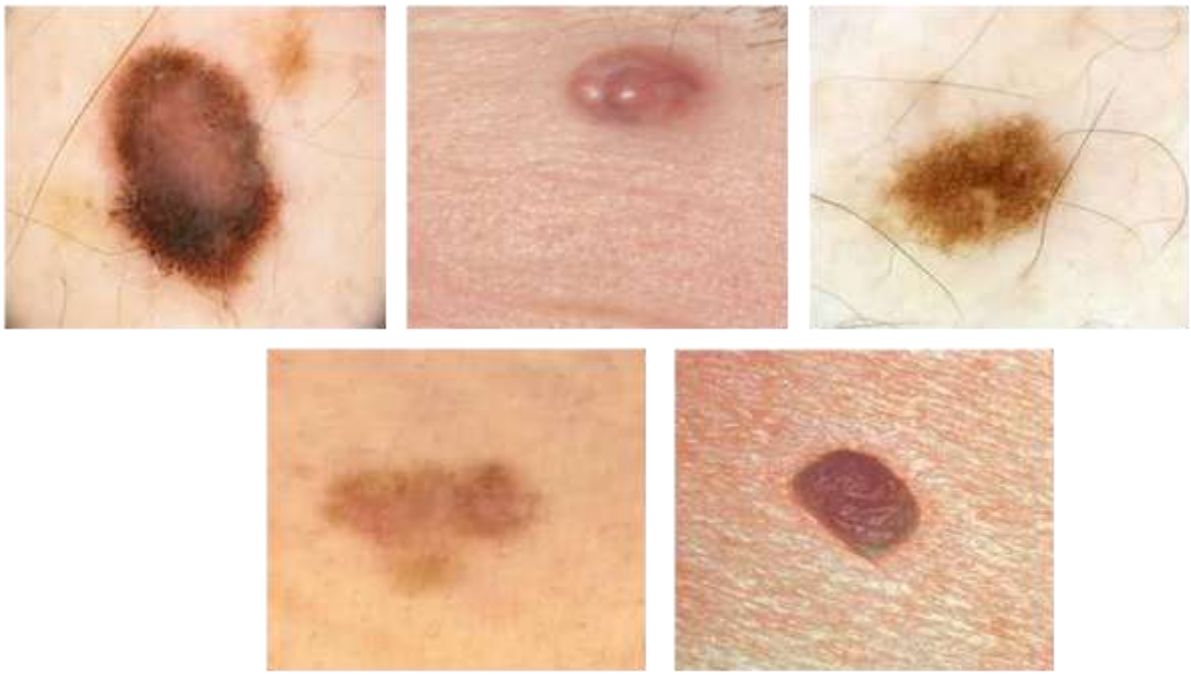

Fig. 2: Input skin images

To evaluate the performance of the classifiers in terms of sensitivity (also called recall in some fields), specificity and accuracy. The formulae for these are given in eqs. (6)-(8). The three terms are defined as follows: Sensitivity (true positive fraction) is the probability that a diagnostic test is positive and it states that the person has the tumor disease, Specificity (true negative fraction) is the probability that a diagnostic test is negative and that the person does not have the disease.

$$Recall\ or\ Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

Accuracy is the probability that a diagnostic test is correctly performed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

A classification with an accuracy of *85%, 96% and 98%* has been obtained by, k-nearest neighbor, support vector machine and convolutional neural networks. The performance of this algorithm is excellent. The application of the proposed method for tracking skin cancer is demonstrated to help pathologists distinguish its type of skin cancer. Table 1 presents the performance of classifiers.

Table 1: Classification Accuracy for the used classifiers

Classifier	KNN	SVM	CNN
Accuracy	89%	97%	98%

CONCLUSIONS

Thus, an automated method for classification of three types of skin cancer is developed based on texture and colour features. This system has been successfully tested on large skin images causing skin

cancer. The proposed system helps the physicians to know about the type of skin cancer, for further treatment. A classification with an accuracy of *85%, 96% and 98%* has been obtained by, k-nearest neighbor, support vector machine and convolutional neural networks. The system can be designed to classify other types of cancers as well with few modifications.

REFERENCES

[1] Ramya, V. J., Navarajan, J., Prathipa, R., & Kumar, L. A. (2015). Detection of melanoma skin cancer using digital camera images. *ARPN Journal of Engineering and Applied Sciences*, *10*(7), 3082-3085.

[2] Mengistu, A. D., & Alemayehu, D. M. (2015). Computer Vision for Skin Cancer Diagnosis and Recognition using RBF and SOM. *International Journal of Image Processing (IJIP)*, *9*(6), 311-319.

[3] Gohila Vani, B., Selvaraj, A., Vijaya Priya, M., & Anandh, N. (2014). Segmentation and Classification of Skin Lesions Based on Texture Features. *International Journal of Engineering Research and Applications, 4*(12): 197-203.

[4] Dua, R., Beetner, D. G., Stoecker, W. V., & Wunsch, D. C. (2004). Detection of basal cell carcinoma using electrical impedance and neural networks. *IEEE Transactions on Biomedical Engineering*, *51*(1), 66-71.

[5] Elgamal, M. (2013). Automatic skin cancer images classification. *IJACSA) International Journal of Advanced Computer Science and Applications*, *4*(3), 287-294.

[6] Ballerini, L., Fisher, R. B., Aldridge, B., & Rees, J. (2012). Non-melanoma skin lesion classification using colour image data in a hierarchical K-NN classifier. In *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, 358-361.

[7] Zarandi, M. F., Zarinbal, M., & Izadi, M. (2011). Systematic image processing for diagnosing brain tumors: A Type-II fuzzy expert system approach. *Applied soft computing*, *11*(1), 285-294.

[8] Haralick, R.M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics, SMC*, *3*(6), 610-621.

[9] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[10] Cristianini, N., & Shawe Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. *First Edition, Cambridge University Press, England.*

[11] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, 588-599.