A Hierarchical Content-Based Image Retrieval Approach to Assisting Decision Support in Clinical Dermatology

M. Nematollahi Arani and H. Ghassemian

Abstract—In this paper, a CBIR tool for supporting clinical decision-making in differential diagnosis of cancerous pigmented skin lesions is proposed. In order to assist dermatologists with a diagnostic decision on dermoscopic image of a difficult dubious pigmented lesion, the proposed CBIR-based CAD system provides them with a set of dermoscopic images of pathologically-confirmed benign or malignant past cases which are of high similarity to the unknown new case in question along with their diagnostic profiles. Relevant past cases are located and retrieved from a comprehensive dermoscopic image set of 1983 benign and malignant skin lesion variations. Corresponding to each image, a 23 dimensional feature vector consisted of the most discriminative shape and color features reflecting the clinical parameters applied by dermatologists for diagnosis of disease, was calculated. Segmentation of lesion images, the very first step to feature extraction, is done via fuzzy C-means color clustering and semi-automatic GVF active contours. Contrary to the conventional CBIR approach which applies only a similarity measurement function to retrieval, here we retrieve similar images through a hierarchical three-step approach. First, SVM classifier classifies the query using the features that best discriminate the classes. Most similar subclass of images within the predicted class in terms of disease severity, stage and type is delineated via FCM clustering in the second step. Finally, similarity measurements are performed within the delineated subclass to retrieve the most similar images to the unknown query image based on the Euclidean distance. Performance evaluation by a physician on 25 query dermoscopic images shows that the proposed dermatologic decision-support CBIR tool improved retrieval precision in terms of critical visual similarity dramatically, indicating high potential of the proposed diagnostic CBIR tool for supporting clinical decision-making in dermatology.

Index Terms—Content-Based image retrieval, hierarchical similarity measurement, computer-aided diagnosis, dermoscopy.

I. INTRODUCTION

Skin cancers, among which Melanoma is the gravest type accounting for nearly four-fifths of the deaths, are most common of all cancers. non-Melanoma skin cancers, namely basal or squamous cell cancers are highly curable if diagnosed and treated early. Despite melanoma’s quick spread to other body parts, it would also be highly curable if diagnosed early and cured properly. Survival rate for melanomas varies between 15 to 65 percent from early to terminal stages whilst only about 80 percent of melanomas are timely diagnosed [1].

Dermoscopy provides dermatologists with a technique for in vivo inspecting of skin lesions, rendering higher accuracy for detecting suspicious cases than it is possible via inspecting with naked eye. It has been calculated that dermoscopy improves the diagnostic accuracy for melanoma by up to 50% depending on the level of experience of the observer in comparison to inspection with the naked eye [2]. The use of digital dermoscopy allows the identification of many different structures and colors invisible to the unaided eye [3].

Recent developments in computer technology and image processing techniques has motivated major research on implementing the concept of computer-aided diagnosis (CAD) for lesion detection and differential diagnosis of different kinds of abnormalities in body parts[4], e.g., abnormalities associated with body skin. Over the last decade, computer aided diagnosis of pigmented skin lesions, whether in the form of classification of malignant and benign lesions or calculating the lesion benignity or malignancy probability, has been an active research area [5]-[6]. In [5], there is a summary table of recent studies on dermoscopic images, including their applied segmentation and classification methods, and comparison of results achieved. In [6], a review of segmentation, feature calculation, feature selection and classification methods applied to the field of computer-aided diagnosis of skin lesions is briefed. However a critical review of the available literature on computer-aided diagnosis of pigmented skin lesions by Rosado and Menzies [4], reveals that the diagnostic accuracy achieved with computer diagnosis was statistically not different from that of human diagnosis(by odds ratios 3.6 vs. 3.51 ; p = 0.80). On the other hand, a study involving 52 volunteer dermatologists investigating the question of how physicians react when faced with a second computerized diagnostic suggestion that contradicts their own diagnoses indicated that physicians do not really welcome a second opinion warmly [7]. In fact, the study showed that only in 24 % of the cases in which the physicians’ diagnoses did not match those of the decision support system, the physicians changed their diagnoses. Also, there was a slight but significant negative correlation between susceptibility to change and experience level of the physicians.

Another approach to Computer-Aided Diagnosis emerged which assumed that the computer output could be utilized by physicians to guide them to a precise diagnosis, but not to suggest them a second diagnosis. In [8], the author believes that providing physicians with a set of
pathologically-confirmed benign or malignant images of past cases which are of high similarity to a new difficult case, can efficiently assist a physician to more confidently render a diagnostic decision. Such a concept can be implemented with a Content-Based Image Retrieval (CBIR) tool for searching and retrieving images most similar to an unknown case, in a database which includes a large number of past cases of known-pathology images similar to those of many new unknown cases. Locating, Retrieving and displaying relevant past cases, provides a physician with a diagnostic support environment rather than a crisp second opinion. In fact, a diagnostic CBIR tool provides a set of visual medical evidences confirming strong probability of presence of one or more diseases. Furthermore, CBIR tools are also used for medical research, medical education, and database management in Picture Archiving and Communication Systems (PACS).

In a typical CBIR, complex information contained in images is fundamentally described by color, shape, texture, and spatial layout descriptors. Similarity matching functions such as Euclidean and Mahalanobis distance functions develop a tool for comparing visual content similarity of images in conventional CBIR systems [9]. Unfortunately, exist literature on CBIR context of skin lesions is not rich. In [10], very simple features have been used. However, the authors take advantage of a data mining algorithm for finding association rules between different feature values. In [11], authors perform aggregation of human perception as a guide in optimizing a visual similarity function in a CBIR system. They only calculated shape features for similarity comparisons. In [12], in addition to providing a fast segmentation algorithms, local color features are calculated by estimating the covariance matrix and histogram analysis of color channels. Moreover, evaluating the statistical parameters of Gray-Level Co-Occurrence Matrix (GLCM) of the lesions, local texture features are also considered. Principal component analysis has been applied to feature space for dimensionality reduction and reducing complexity, in [11]-[12]. In [13], making use of a combination rule based on the Dempster-Shafer theory, a neural network classifier is used in collaboration with a CBIR tool. In [14], we calculated most relevant features of shape, texture and color corresponding to dermoscopic criteria so as to improve retrieval accuracy. Performance evaluation of all of these dermoscopic CBIR tools indicate the potential of CBIR concept for supporting clinical decision making in dermatology.

However, our investigations of previous dermatological CBIR systems indicate that they seriously suffer from the downsides of the so-called semantic gap (e.g. a difference in concepts between humans and machines) due to application of the traditional approach to similarity measurement. As previously mentioned, in the traditional approach to CBIR each image is represented by a vector of feature values in the image database and images that are most similar to the query in terms of some distance measure (e.g., Euclidean distance) are retrieved. For this approach, the choice of similarity metric and features to include for characterizing the images are critical factors in its ability to achieve high retrieval precision [15]. Moreover, the relevance between the query and target images is in proportion to feature similarities without any semantic interpretation of images. Images with high feature similarities to the query image may be very different from the query image in terms of the semantics perceived by the user or semantics according to the predefined categories. This is referred to as the well-known semantic gap problem which reflects the discrepancy between the relatively limited descriptive power of low-level image features and the high-level user semantics and hinders the efficient judgment of similarity among images through calculation of feature vectors only [9].

In order to narrow down the semantic gap in retrieval process, different strategies based on the level of user involvement have been applied in CBIR literature [16]. Semantic image categorization based on machine learning techniques such as classification and clustering is a promising approach towards effective retrieval in Medical Image Databases in which images are semantically organized in not only different modalities such as CT, MRI, PET, dermoscopy and etc. but also different body organs [17]. Furthermore, images of a particular body organ in a specific modality are usually semantically organized in different disease categories. In fact, although defining similarity is a difficult task because of the difficulty of human perception modeling, in medical domain similarity means images that correspond to the same disease type, stage, severity, and treatment [15].

In the field of Dermoscopic images of pigmented skin lesions, lesion images also fall into diverse categories based on various disease type, stage and severity. In the very first phase, pigmented lesions are usually categorized as benign or malignant melanoma. However, two lesions belonging to the same category might be so different in lesion type or disease stage that they illustrate noticeably diverse appearances. That is, there are sub-classes in each major category of benign or malignant. For example, each of the benign lesions may represent a Basal Cell Carcinoma, Dermatofibroma or Dysplastic Nevi. Malignant lesions also may appear very differently depending on disease severity or stage at the very least [3]. Therefore, regarding this property of dermoscopic lesions and images, we applied machine learning techniques in a hierarchical approach to similarity measurement in order to deal with the semantic gap and enhance visual similarity of retrieved images in our dermatological CBIR system.

The hierarchical approach is implemented in three steps. First, a classifier determines whether the new query image is a benign or malignant lesion. Secondly, since database images are only labeled in two classes of benign or malignant lesions and labels of images belonging to sub-categories of these classes based on different disease type or stage are not available to us, a clustering algorithm in the second step determines to which sub-class within a disease class a typical query belongs. Ultimately, the most similar images to the query image are retrieved based on the Euclidean Distance measure from among images of the chosen sub-class within the predicted disease class. Furthermore, since a hierarchy of classes exists and not all pairs of images within one class have equivalent perceptual similarity, the features that most effectively discriminate among images from different classes may not act the most
effectively for retrieval of images belonging to the same subclass within a class [15]. Therefore, the classification step will be implemented using the features that best discriminate among benign and malignant images and moreover, the clustering step will be implemented using the features that best discriminate the subclasses within each category. That is, feature selection algorithms will be wrapped around the classification and clustering algorithms in the first and second steps respectively so that in each of these steps a minimum set of most discriminative features are selected for classification and clustering implementations.

The rest of the paper is organized as follows. Section II describes the dataset collection applied in this study. Section III explains the lesion segmentation phase of the proposed CBIR system. Feature extraction is discussed in Section IV. Furthermore in Sections V and VI, similarity measurement analyses are presented. Performance evaluation of the proposed hierarchical dermatological CBIR system in comparison with its traditional counterpart is elaborated in Section VII. Finally, conclusion and future work are given in Section VIII.

II. DATASET DESCRIPTION

A very important necessity of a CBIR system for helping clinical decision-making is the availability of a comprehensive dataset including images of a large number of past pathologically-confirmed cases from different kinds of diseases threatening a particular organ which are supposed to be of high similarity to those of many new unknown cases [3].

In this study a collection of 1983 dermoscopic images of pigmented skin lesions in two categories of benign and melanoma lesions (acquired from Dermatology Department of University of Graz, Austria) are utilized. This collection is consisted of dermoscopic images of 126 melanomas and 1857 benign lesions which have been taken with MoleMax II system in 640×480 resolutions. However, there are no validated labels available within benign and melanoma categories in terms of disease type and severity.

III. LESION SEGMENTATION

Lesion segmentation is the very first step to feature extraction in computerized analysis of skin lesions. Since transition from lesion interior to healthy skin in dermoscopic images is very smooth, lesion detection and determining lesion border is a difficult task in this field. However, many studies have circumvented this challenge incorporating several segmentation algorithms or via combining segmentation and border detection algorithms [6], [18]-[19]. Imposed complexity is justified by extensive dependence of accurate shape description on lesion boundary delineation results.

Prior to segmentation some image pre-processing analysis was necessary the most critical of which was hair removal of dermoscopic images. Since hairs covering the skin lesion can have disturbing influence on the derivation of the correct boundaries, Dullrazor software was employed in order to remove hair off the lesion image surface [20]. However, Hair pixels on the lesion dermoscopic image constitute a very trivial percentage of the total image pixels.

Fig. 1. Presence of hair on lesion surface influences lesion segmentation.

So in further steps of feature calculation, the information is taken again from the original lesion image, but now with better defined lesion boundaries.

Color conveys significant information in dermoscopic images since lesion regions have a colorimetric signature, i.e. they are completely defined by their color and a color-based segmentation scheme seems to be suitable as a first segmentation step. Clustering techniques should be particularly suited since the spatial information is not used at this stage [19]. Color clustering is intended to discriminate interior lesion pixels from healthy skin pixels. It is noteworthy that dermoscopic images of pigmented lesions show different shades in color and there rarely is a hard transition between neighboring regions. That is, transition from lesion interior to surrounding skin is very smooth and fuzzy. Therefore, we applied Fuzzy C-Means (FCM) color clustering algorithm in RGB color space [21]. Number of clusters for the FCM algorithm to settle in was initially set equal to two, regarding the lesion region and surrounding healthy skin. The clustering process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. Using MATLAB Fuzzy Logic toolbox, maximum number of iterations for the FCM algorithm was set to the average value of 80 with regard to both operation run time and validation of the segmentation results of 25 randomly-selected cases through visual inspection by a physician. Minimum amount of objective function improvement was set to default value of MATLAB software for this parameter ($10^{-5}$). Finally, FCM color clustering segmentation results were assessed visually one by one. In 91% of the cases, the segmentation results were considered as satisfactory by a physician.

Since most segmentation breakdowns occurred with Melanoma lesions which constituted the minority class, though of great importance, of our dermoscopic image collection, Gradient Vector Flow active contour software was semi-automatically applied as the second step to lesion segmentation in order to compensate for the shortcomings of FCM color clustering of dermoscopic images [22]. The initial snake in GVF algorithm was set equal with the rough result of the segmentation by FCM color clustering. One example of lesion segmentation with regard to hair removal influence is illustrated in Fig. 1.
IV. FEATURE EXTRACTION

In CBIR, as previously stated, visual contents of images in the form of multi-dimensional feature vectors are calculated so that each image is represented by a vector of feature values in the image database. Feature vectors are used to search the similar cases of a query image. In content-based access to medical images for supporting clinical decision-making, the features calculated and included in vectors representing images in the dataset need to be a reflection of clinical parameters applied by physicians for diagnosis of disease. Diagnostic algorithms most common in differential diagnosis of pigmented skin lesions in routine clinical work are pattern analysis, the ABCD rule of dermoscopy, the 7-point checklist, Menzies method, and revised pattern analysis. One can refer to [3] for detailed parameters regarded each algorithm. Generally, global features of shape, texture and color are included in these algorithms. In order to perform texture feature analysis, lesion regions of all dataset images will be required to be explored by a dermatologist so that lesion patches with specific texture structures of diagnostic value be professionally delineated. Using texture analyses, one should take into account that discriminatory power of texture features will be effective and will improve results significantly if only dataset images are of high quality and fine textures of lesions’ surface is not obscured in poor image resolutions; otherwise the retrieval accuracy might be hindered by inaccurate textural characteristics of lesion images. It is noteworthy that overall characteristics of shape and color in dermoscopic images are not as vulnerable to loss as their fine textural structure. Since in a significant portion of our database images, fine textures of lesions’ surface were vague, we failed to analyze texture features in our CBIR system. Moreover, local features such as particular dermoscopic structures are important in evaluation of skin lesions. However, extraction of these structures has also been proved to be a difficult task and very susceptible to image acquisition conditions. Hence, in this work we concentrate on calculating most relevant global features of shape and color.

A. Feature Calculation

Applied shape features have to be robust to spatial transformations. That is to say, they should be invariant to translation, rotation, scaling and transformation of starting point of contour representation sequence. They also need to be computationally efficient and inexpensive. Here in this study we applied 14 simple shape features such as area, eccentricity, perimeter, convex perimeter, elongation, compactness to name but a few [23]. Simple shape features are of low computational cost but they result in too many false positives, which make their sole use inefficient. More efficient methods are Moment Invariants (MIs) and Fourier Descriptors. Moment-based shape descriptors are used when a region-based analysis of the object is performed. Region-based analysis exploits both boundary and interior pixels of an object. These shape descriptors are more robust to noise and distortions. Moment is popular for region-based analysis as central moments are invariant to translation, rotation and scale. They are also computationally simple. Moment analysis describes essential and frequently used shape features [24].

Moreover, Fourier descriptors elaborate the boundary regardless of its interior more precisely. Spatial transformations are directly compensated for through taking advantage of Fourier Transform properties. Representing the contour of a shape with a sequence of complex numbers in Cartesian coordinate’s plane, Fourier descriptors describe shape of an object by the Fourier transform of its boundary [25]. From Fourier characteristics we know that lower frequency components are description of general shape of an object, whereas higher frequency components represent shape’s fine details. Regarding this concept, border irregularity and asymmetry can be analyzed very simply. Here in this work, Fourier shape feature vector for each image includes first 10 harmonic coefficients of Fourier transform which form a 40 dimensional vector since the correlation coefficient between original images and reconstructed images using these 40 coefficients was calculated 98% in average. However, it is possible to improve retrieval performance by applying more coefficients if necessary. Therefore, a 61-dimensional shape feature vector consisted of 14 simple shape features, 7 Moment Invariants and 40 Fourier features were calculated for lesion boundary of each dermoscopic image in the database.

Many color spaces have been designed to facilitate the color specification and the key issue always being considered in selection of a perceptually uniform color space. HVC color space comes from Munsell color coordinate system, which is considered for its successful imitation of human color perception. Local color feature in the form of a feature vector is extracted by considering the mean or average color of the lesion in HVC color space and variance-covariances of the color channels by estimating the covariance matrix [26]. Since the covariance matrix is symmetric, only 6 values of it need to be stored in the feature vector for later similarity matching. These variance-covariance values along with the color channels mean vector constitute a 9 dimensional color feature vector.

B. Feature Normalization

Feature normalization is an important preprocessing step that is necessary to prevent features with large ranges from dominating the calculations and also to avoid numerical instabilities [5]. One of the most common normalization methods is the Linear Scaling to Unit Variance in which the feature component is transformed to a random variable with zero mean and unit variance as \( x' = (x - \mu) / \sigma \). Where \( \mu \) and \( \sigma \) is the sample mean and the sample variance of that feature respectively. Under the normality assumption, an additional shift and rescaling as: \( x' = (((x - \mu) / 3\sigma) + 1)/2 \) guarantees 99% of \( x' \) to be in the \([0, 1]\) range. We can then round off the out-of-range components to either 0 or 1 [27].

C. Feature Selection

The curse of dimensionality is a term coined by Bellman in 1961 and refers to the problems associated with multivariate data analysis as the dimensionality of feature space increases. In practice, the curse of dimensionality
There are two general approaches for performing dimensionality reduction: feature extraction and Feature selection. Feature extraction approach the examples of which are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) transforms the existing features into a lower dimensional space. In this method, a large number of features are evaluated in the test bed and only a few are selected for the final implementation. So when features are expensive to obtain this approach will not be applicable. In addition, when you transform or project, the measurement units of your features (length, weight, etc.) are lost and extraction of meaningful rules from a typical classifier will not be possible. But feature selection approach, also called feature subset selection, selects a subset of the existing features without a transformation. Feature Subset Selection requires a search strategy to select candidate subsets and an objective function to evaluate the goodness of these candidates. In fact, objective function is a feedback signal used by the search strategy to select new candidates. Objective functions are divided in two groups: filters in which the objective function evaluates feature subsets by their information content, typically interclass distance, statistical dependence or information-theoretic measures, and wrappers in which the objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical re-sampling or cross-validation [28]. In this study, we use both filter and wrapper approaches. As previously elaborated, a 61 dimensional shape feature vector was calculated for every image in the dermoscopic dataset in comparison with a 9 dimensional color feature vector. We applied the filter feature selection methodology to 61 dimensional shape feature vectors in order to keep the number of shape features balanced with the number of color features. We chose \( t \)-test (Absolute value two-sample \( t \)-test with pooled variance estimate) and \( \text{roc} \) (Area between the empirical receiver operating characteristic (ROC) curve and the random classifier slope) criteria so as to assess the significance of every feature for separating two labeled groups (benign nevi and melanoma). Among the total number of 61 shape features, 14 features which were considered as most discriminative by both aforementioned criteria were selected and substituted the original shape feature vectors. Accumulation of the 14 shape features and 9 color features resulted in a 23 dimensional feature vector for every image in the dataset.

Moreover, wrapper feature selection approach was applied along with classification and clustering phases. As it were previously mentioned, features that are most effective in discriminating among images from different classes may not be the most discriminative for identification of images belonging to diverse subclasses within a class. Therefore, in each of dataset classification and within-class clustering phases, a wrapper approach will be implemented to obtain the best retrieval result in terms of enhanced visual similarity and reduced semantic gap.

Since there are 23 features available and feature subset selection will involve \( 2^{23} \) combinations, exhaustive search is unfeasible. We employed Sequential Forward Selection (SFS) in our implementation to search the features. Other search methods in this regard are Sequential Backward Selection (SBS), Plus-L Minus-R Selection (LRS), Bidirectional Search (BDS), Sequential Floating Selection (SFFS and SFBS) and Genetic Algorithms [29]. Sequential Forward Selection is the simplest greedy search algorithm that adds one feature at a time. SFS performs best when the optimal subset has a small number of features and its main disadvantage of SFS is that it is unable to remove features that become obsolete after the addition of other features [30]. Although, SFS does not provide us with an optimal solution, but it is a simple search strategy with \( O(d^2) \) complexity and is sufficient for our purpose.

V. SIMILARITY MEASUREMENT

As previously mentioned, each image in the traditional approach to CBIR is represented by a vector of feature values in the image database and images that are most similar to the query in terms of some distance measure (e.g., Euclidean distance) are retrieved. Block diagram of the traditional approach to CBIR is illustrated in Fig. 2. It was also further explained that these systems suffer from the downsides of the so-called semantic gap e.g. dramatically diverse appearances of retrieved images. In order to narrow down the semantic gap in retrieval process, a hierarchical approach to retrieval of dermoscopic images is applied in this study the block diagram of which is presented in Fig. 3. The hierarchical similarity measurement approach is
implemented in three steps. First, a classifier determines whether the new query image is a benign or malignant lesion. Secondly, a clustering algorithm delineates to which sub-class within a disease class a typical query belongs. Ultimately, the most similar images to the query image are retrieved based on the Euclidean Distance measure from among images of the chosen sub-class within the predicted disease class.

VI. SIMILARITY MEASUREMENT EXPERIMENTS

A. SVM Classification

Classification experiments are described in this section. A review of exist literature in the field of classification of pigmented skin lesions indicates that SVM classification has resulted in most accurate predictions in contrast to other classification techniques [5]. Hence, we also utilize SVM classification in order to isolate benign lesions from malignant melanomas in the first step of hierarchical similarity measurement.

Detailed information on theoretical background of Support Vector Machines can be found in [31]-[32]. We used the STPR pattern recognition toolbox to perform SVM classification [33]. The radial basis function (RBF) kernel was adopted since linear kernel cannot handle non-linearly separable classification tasks and is a special case of the RBF kernel, the computation of the RBF kernel is more stable than that of the polynomial kernel which introduces values of zero or infinity in certain cases, polynomial kernel has more parameters needed to be determined, sigmoid kernel may not be positive semi-definite and for certain parameters it behaves like RBF [32].

1) Initial Experiments with the SVM Classifier

When using RBF kernel, values of kernel parameters, C (cost/penalty) and \( \varphi \) (kernel width), need to be appropriately determined so that the classifier be able to accurately predict unknown (test) data. Using 10-fold cross-validation in order to evaluate the goodness of a particular combination of parameter values [34], a grid search on two parameters of C and \( \varphi \) in exponentially growing sequences of values, \( C \in \{2^1, 2^2, ..., 2^8\} \) and \( \varphi \in \{21.5, 21.13, ..., 23\} \), was performed. Performing the grid search procedure on dataset of images with 1983 dermoscopic images, pairs of \( C \) and \( \varphi \) were tried and the one with the best cross-validation accuracy was picked [32], [5]. Optimum values of \( C^* \) and \( \varphi^* \) equal with 8 and 0.5 respectively resulted in maximum values of sensitivity and specificity equal with 30% and 96.24% respectively. Authors in [5] believe that the unsatisfactory result for sensitivity is due to class distribution imbalance. Since our dataset also holds same class distribution imbalance (1857 benign nevi: 126 melanomas), we particularly deal with it in next section.

2) Dealing with Class Distribution Imbalance

Class distribution imbalance happens when some classes in a typical dataset dramatically outnumbers other classes. When classes are not approximately equally represented, most classifiers focus on learning the majority class and therefore classification accuracy for minority classes will be poor. In domains, such as medical diagnosis, the misclassification costs are often unequal and classifying the minority (malignant) samples as majority (benign) leads to catastrophic consequences [35].

Dealing with class imbalance has been an interesting problem in machine learning community. Reviewing the existing literature in this regard, authors in [35] showed that a combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than other proposed strategies such as sole under-sampling of the majority class and varying the loss ratios in Ripper or class priors in Naive Bayes. Under-sampling refers to random elimination of majority class samples and over-sampling involves creation of synthetic minority class examples. According to SMOTE over-sampling algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated in the following way. The difference between the feature vector (sample) under consideration and its nearest neighbor is calculated. Multiplying this difference by a random number between 0 and 1, and adding it to the feature vector under consideration, a random point along the line segment between two specific features is selected. This approach effectively forces the decision region of the minority class to become more general [35].

In our current implementation, 5 neighbors are randomly chosen among \( k = 10 \) nearest neighbors so that minority melanoma class grew in population from 126 to 630. On the contrary, through under-sampling majority class members were randomly removed from the database resulting in a 476:440 benign to melanoma ratio.

Furthermore, in order to find the features that most effectively discriminate among images from different classes SFS feature selection algorithm was wrapped around the SVM classifier with optimum values of \( C^* \) and \( \varphi^* \) equal with 8 and 0.5. In an iterative procedure applied to the sampled dataset (476:440 benign to melanoma), starting from the empty set, SFS sequentially adds the feature \( x^* \) that results in the highest improvement to our feature selection criterion, \( J(Y^* + x^*) \), when combined with the features \( Y^* \) that have already been selected [30].

Here, we chose sensitivity and specificity criteria as feature selection criteria the values of which are calculated with 10-fold cross-validation in each iteration. In each step of the iterative procedure, the feature which results in maximum improvement in sensitivity and specificity values is added to the selected feature subset. Iterative SFS will come to a halt when adding more features cause no more improvement in sensitivity and specificity values. The iterative procedure performed in the classification step is as follows:

1. Start with an initial empty set of features: \( Y_0 = \{\phi\} \)
2. Select the next best feature: 
   \[ x^* = \arg\max_{x \not\in Y^*} \{ J(Y^* + x) \}. \]
   Sensitivity and specificity values, for performance evaluation of SVM classification, are considered as feature selection criteria (objective function) the values of which are calculated via 10-fold cross-validation.
pass algorithm for estimating the number of clusters and the
Subtractive Clustering and applied for initialization of FCM
clusters and initial cluster centers are calculated through
for a selected subset of features, the optimum number of
categories. During every step of these iterative procedures,
find a minimum set of features that best discriminate the
FCM clustering algorithm in an iterative procedure so as to
clusters are of important consideration.

most discriminative features for most effective natural
performed for benign and malignant classes for each of
clustering. Here in this study, clustering is separately
toolbox of MATLAB software in order to perform FCM
be found in [36], [37]. Here, we used the Fuzzy logic
theoretical background of FCM clustering algorithm can
utilize the Fuzzy C-Means (FCM) which is the most widely
clusters to each input pattern (feature vector). Here we
problem, which assigns degrees of membership in several

B. FCM Clustering

In the second phase of the hierarchical similarity measurement phase of diagnostic dermatological CBIR system, a clustering algorithm delineates to which sub-class within a disease class a typical query belongs. Clustering aims to find the natural groupings exist in the feature space of the images of each class. Fuzzy clustering models have proved a particularly promising solution to the clustering problem, which assigns degrees of membership in several clusters to each input pattern (feature vector). Here we utilize the Fuzzy C-Means (FCM) which is the most widely used fuzzy clustering algorithm. Detailed information about the theoretical background of FCM clustering algorithm can be found in [36], [37]. Here, we used the Fuzzy logic toolbox of MATLAB software in order to perform FCM clustering. Here in this study, clustering is separately performed for benign and malignant classes for each of which finding the optimum number of clusters, selecting the most discriminative features for most effective natural groupings and evaluating the goodness of discovered clusters are of important consideration.

We use SFS feature selection algorithm wrapped around FCM clustering algorithm in an iterative procedure so as to find a minimum set of features that best discriminate the subclasses in each of the benign and malignant melanoma categories. During every step of these iterative procedures, for a selected subset of features, the optimum number of clusters and initial cluster centers are calculated through Subtractive Clustering and applied for initialization of FCM clustering algorithm. Subtractive clustering is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data [37]. The subtractive clustering method assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the density of surrounding data points. The algorithm first selects the data point with the highest potential to be the first cluster center and removes all data points in the vicinity of the first cluster center (in a limited adjustable radius) in order to determine the next data cluster and its center location. This process is iterated on until all of the data is within the limited radii of the cluster centers [37]. The value of radii for clustering of benign and melanoma classes were adjusted equal with 0.08 and 0.3 respectively.

The cluster estimates, number of clusters and initial cluster centers, are used to initialize FCM clustering in each step of iterative procedure. FCM assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point’s membership grade. Having performed the iterative FCM clustering in each step of iterative procedure, we evaluate the clusters discovered by our candidate feature subset with the trace(S−1Sb) criterion (the objective function), a frequently used criterion in discriminant analysis, in order to measure how well the candidate feature subset isolates the clusters (i.e., subclasses). Where, Sb measures how scattered the samples are from their cluster means (compactness) and Sz measures how scattered the cluster means are from the total mean (separability) [38]. We would like the distance between each pair of samples in a particular cluster to be as small as possible and the cluster means to be as far apart as possible with respect to the chosen similarity metric. Sb and Sz are defined as follows:

\[ S_b = \sum_{i=1}^{k} \mu_i \sum_{j=1}^{k} \pi_i (X - \mu_i)^T (X - \mu_i) / w_i \]

\[ S_z = \sum_{i=1}^{k} \pi_i (\mu_i - M_0)^T (\mu_i - M_0) \]

\[ M_0 = E(X) = \sum_{i=1}^{k} \pi_i \mu_i \]

where, \( \pi_i \) is the probability of class \( w_i \). \( X \) is a random
feature vector representing the image, $\mu_i$ is the mean vector of class $w_i$, $M_i$ is the total mean across all data points or images in the database, $w_i$ is the class $w_i$, $\Sigma_i$ is the covariance matrix of class $w_i$, and $E[.]$ is the expected value operator. Among the many possible separability criteria, we chose $\text{trace}(\Sigma_i^{-1}S_i)$ our criterion because it is invariant under any nonsingular linear transformation [15].

In each step of the iterative procedure, the feature which results in maximum improvement in $\text{trace}(\Sigma_i^{-1}S_i)$ value is added to the selected feature subset. Iterative SFS will come to a halt when adding more features cause no more improvement in this criterion. The iterative procedure performed in the clustering phase is as follows.

1. Start with the empty set: $Y_i = \{\phi\}$
2. Select the next best feature: $x^* = \text{arg max}_{x \in S_i} \{\text{trace}(\Sigma_i^{-1}S_x)\}$. The value of $\text{trace}(\Sigma_i^{-1}S_x)$ is considered as feature selection criterion for performance evaluation of FCM clustering.
3. Update: $Y_i = Y_i + x^*$; $k = k + 1$
4. If $(Y_i) \geq J(Y_{i-1})$, go to 2.
5. If $(Y_{i-1}) \leq J(Y_i)$, then $Y_k = Y_i$.
6. End.

Detailed results of the Iterative Sequential Forward feature Selection algorithm for FCM clustering of benign and melanoma classes are in Table II and Table III. According to Table II, maximum value of $\text{trace}(\Sigma_i^{-1}S_x)$ equal with 11.67 is yielded for a 3-member feature subset which has resulted in 3 subclasses for melanoma class. According to Table III, maximum value of $\text{trace}(\Sigma_i^{-1}S_x)$ equal with 573.46 is yielded for a 4-member feature subset which has resulted in 20 subclasses for benign nevi class. However, since some of the discovered clusters were too sparse with 20 clusters, we stepped back one iteration and accepted 15 subclasses discovered by a 3-member feature subset for benign nevi class. Hence, FCM clustering will be used to determine the unknown query image’s subclass in its specific predetermined class.

C. Retrieval

Submitting a new unknown query image to dermological CBIR system, visual content of the query image is described in a feature extraction phase. A multidimensional feature vector representing the query in dataset’s feature space is produced. Moreover, in the first step of similarity measurement phase, benignity or malignancy of the query image is determined by SVM classifier. Having known to which disease class the query belongs, Mahalanobis distance of the query feature vector from the feature vectors of cluster centers of the distinguished class are calculated. From among all existing clusters in the determined class, three of the clusters minimally distanced from the query feature vector are selected. Ultimately, the most similar images to the query image, those with the minimum distance from the query image based on Euclidean distance measure, are retrieved from among images of the three chosen sub-class.

VII. PERFORMANCE EVALUATION

In order to evaluate the performance of the proposed system, a collection of 1983 dermoscopic images of pigmented skin lesions in two categories of benign and melanoma lesions are utilized. This collection is consisted of dermoscopic images of 126 melanomas and 1857 benign lesions. However, there are no validated labels available within benign and melanoma categories in terms of disease type and severity. Regarding the difficulty of collecting quality homogeneous dermoscopic images, our dataset of 1983 dermoscopic images is large and generalized enough in comparison with the number of images in other studies dermatological CBIR (184, 358 and 200 in [23], [12] and [13] respectively). Two methods of similarity measurement dermatological diagnostic CBIR system are compared in this section. Retrieval performance of three-step hierarchical similarity measurement is compared with Retrieval performance of traditional similarity measurement based on Euclidean distance measure.

A. The Method

A physician was asked to evaluate the retrieval results. We randomly selected and removed 25 dermoscopic images (7 melanomas and 18 benign nevi regarding the original class distribution) from the dataset in order to use them as query test images. For every query image, 9 images most similar to the query from the database were retrieved and showed to the physician. During test conduction, the physician was not aware of the difference in retrieval methods. Performance evaluation was performed two times for each query image regarding the two similarity measurement techniques. Physician compared the query results...
image with its 9 retrieved similar images and assigned to each of retrieved images a score in the range of 0 to 10 based on the level of visual similarity to the query image. The score 10 for a retrieved image meant completely similar in terms of shape and color characteristics and a zero score meant totally different. The physician compared to typical retrieved lesions in terms of elongation, roundness, irregularity of lesion border, size and diameter ratios to explore lesions’ shape and also in terms of color characteristics by tracking the presence of patches of diagnostically discriminative colors such as blue-white, dark and white and yellow in the two lesions.

In order to compare the performance of two retrieval methods, retrieval precision was calculated for two strategies. Precision is defined as the number of correct retrievals divided by the total number of images retrieved. In our experiments, we considered those images with a similarity score of 5 to 10 to be correct retrievals and those images scoring lower than 5 to be irrelevant cases. Recall is another criteria applied to performance evaluation of retrieval systems. However, since labels of images in subclasses of benign and melanoma categories are not available to us, we could not calculate this criterion.

B. Assessment Results

Performance evaluation results are presented in Table IV. For hierarchical similarity measurement technique, 174 of the total 225 retrievals were assigned scores more than 5 by the physician and retrieval precision was calculated equal with 77.33%. However, Considering the SVM classification error 96.67%, retrieval precision is 74.75% for this method. On the other hand, traditional similarity measurement approach based on Euclidean distance measure yielded 118 correct retrievals resulting in retrieval precision of 52.44%. Performance evaluation was conducted based on visual assessment of retrieval results. Comparison of retrieval results of two methods indicates that hierarchical similarity measurement outperforms the traditional approach considerably (by about 25%) in terms of visual similarity and diminishing of the semantic gap.

VIII. Conclusions

In this work, A CBIR approach to computer-aided diagnosis was proposed so as to assist physicians in diagnosis of cancerous pigmented skin lesions. Providing physicians with a set of pathologically-confirmed benign or malignant images of past cases which are of high similarity to a new difficult case, can help them to more confidently render a precise diagnostic decision and the CBIR software system implemented regarding this concept performs as a powerful instructive tool for inexperienced physicians. However, few investigations of previous dermatological CBIR systems indicate that they seriously suffer from the downsides of the so-called semantic gap problem resulting in dramatic visual dissimilarities of retrieved images.

Here in this study, we applied machine learning techniques in a three-step hierarchical approach to similarity measurement in order to narrow down the semantic gap in retrieval process and enhance visual similarity of retrieved images in our CBIR system.

Performance evaluation of the traditional dermatological CBIR in comparison with the hierarchical approach indicates that visual dissimilarities have been faded considerably through application of the latter approach. Although the retrieved images were chosen from among the images in the same disease class (as determined by the SVM classifier), there were many cases for which the physician assigned a similarity score of near zero because they had different disease severity, structure, and/or visual appearance. This happened less frequently in the retrieval results of the hierarchical approach. We conclude that since lesion images fall into diverse categories based on various disease-stages and severity and visual appearance, retrieval of images based on just disease class is not sufficient and effective to assist physicians with dermatological diagnosis. In diagnostic dermatological CBIR systems, retrievals should be performed on the basis of visual similarity so that such a system is applicable to diagnostic tasks or instructive purposes.

It is notable that exist literature on CBIR context of skin lesions is not very rich and yet none of the a few present studies care for the inter-class diversities exist among pigmented skin lesions. On the other hand, a precise comparison of our study with the existed literature demands the dermoscopic image dataset applied to be unique. However, the very fundamental barrier which generally hinders study and development of CBIR approaches is the lack of a unique reliable test-bed and standard measurement criterion for comparison of different studies.

There are several possibilities in order to improve retrieval accuracy of the proposed system. Analysis of texture features, a very chief diagnostic clue in clinical dermatology, has been ignored in this study. Since texture diversities is a commonplace diagnostic clue and cause substantial visual dissimilarities in dermoscopic images, their exploration in diagnostic CBIR will be a promising step towards defeating the semantic gap. However, using texture analyses, one should take into account that discriminatory power of texture features will be effective and will improve results significantly if only dataset images are of high quality and fine textures of lesions’ surface is not obscured in poor image resolutions; otherwise the retrieval accuracy might be hindered by inaccurate textural characteristics of lesion images. It is noteworthy that overall characteristics of shape and color in dermoscopic images are not as vulnerable as their fine textural structure. Furthermore, so as to enhance retrieval performance of the proposed system, structural analysis such as detection of blue-white veil, asymmetric blotches and pigment networks also not only are of great diagnostic importance to clinical dermatology, but also they are the critical key to Region-based Retrieval, a retrieval strategy with the aim of
narrowing down the semantic gap. Relevance feedback (RF) and adaptive similarity matching functions are of other popular concepts which aim to narrow down the semantic gap and improve image understanding and retrieval based on some degree of user involvement in the retrieval process.

ACKNOWLEDGMENT

We would like to thank Dr. Harald Kittler for providing us the comprehensive collection of dermoscopic images. We would also like to thank Dr. Leila Nematollahi for her valuable help in evaluation experiments.

REFERENCES


Maryam Nematollahi Arani was born in Aran-$\&$-Bidgol, Iran, in 1983. She received her B.Sc. degree in electrical engineering from Shahrood University of Technology, Shahrood, Iran, in 2005 and her M.Sc. degree in biomedical engineering from Tarbiat Modares University, Tehran, Iran, 2009. Since 2009, she is working with Beena Vision Systems Company as an image processing algorithm developer. Her research interests include theory and applications of computer vision, content-based image retrieval and statistical pattern recognition in industry and medical field.

H. Ghassemian was born in Iran in 1956. He received the B.S.E.E. degree from Tehran College of Telecommunication in 1980 and the M.S.E.E. and Ph.D. degree from Purdue University, West Lafayette, USA in 1984 and 1988 respectively. Since 1988, he has been with Tarbiat
Modares University in Tehran, Iran, where he is a Professor of Computer and Electrical Engineering.

His research interests focus on Multi-Source Signal/Image Processing and Information Analysis and Remote Sensing. Dr. Ghassemian has published more than 200 articles in peer-reviewed journals and conference papers. He has trained more than 80 M.Sc. and Ph.D. students who have assumed key positions in software and computer systems design applications related to signal and image processing, in the past 20 years. He is an IEEE Senior member.