Expert Systems with Applications 42 (2015) 6578–6585

Contents lists available at ScienceDirect

Expert Systems with Applications

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

MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images

Ioannis Giotis [a],*, Nynke Molders [b], Sander Land [c], Michael Biehl [a], Marcel F. Jonkman [b], Nicolai Petkov [a]

[a] *Johann Bernoulli Institute for Mathematics and Computing Science, University of Groningen, Groningen, The Netherlands*
[b] *Department of Dermatology, Center for Blistering Diseases, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[c] *Department of Biomedical Engineering, Kings College London, United Kingdom*

ARTICLE INFO

Article history:
Available online 1 May 2015

Keywords:
Computer-assisted diagnosis
Classification
Dermatology
Medical imaging
Melanoma

ABSTRACT

Melanoma is one of the most aggressive types of skin cancer and in many cases it is difficult to differentiate from benign naevi. In this contribution we present a decision support (expert) system, which we call MED-NODE, able to assist physicians with this challenging task. The proposed system makes use of non-dermoscopic digital images of lesions from which it automatically extracts the lesion regions and then computes descriptors regarding the color and texture. In addition, a set of visual attributes is provided by the examining physician. The automatically extracted descriptors and the attributes provided by the physician are separately used for automatic prediction. Final classification is achieved by a majority vote of all predictions. The proposed system achieves high diagnostic accuracy results (81%) and performs comparably to state-of-the-art methods that are using dermoscopic images, though such images contain more detailed information and are subject to less noise and illumination effects. The simple input requirements and the robustness of its descriptors allow MED-NODE to be an effective tool within the diagnostic process for melanoma. In addition, the modular nature of the system allows for it to be easily extended.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The incidence of melanoma is increasing by 3–7% in the fair-skinned population globally, affecting approximately 40 of 100,000 persons per year (Marks, 2000). In 1985 the ABCD diagnostic acronym (Friedman, Rigel, & Kopf, 1985) for melanoma was devised in response to the increase of melanoma incidents in the United States. The initials stand for the main diagnostic criteria used to recognize melanoma: Asymmetry, Border irregularity, Color variation and a Diameter greater than 6 mm. Differential diagnosis of melanoma, in particular from benign melanocytic naevi is not straightforward. Hence, a growing interest has developed in the automated analysis of digital images obtained by epiluminescence microscopy (dermoscopy) in order to assist dermatologists in this task.

A study comparing the performance of computer diagnostic systems against those of trained physicians was presented in Rosado et al. (2003) reporting mixed results. Computer systems generally achieved higher sensitivity but lower specificity, yielding more false positives than humans. Another study (Ganster et al., 2001) was conducted in 2001 using a total of 5363 dermoscopic images. In that study the lesions are segmented using a fusion of the results of different segmentation techniques, and subsequently a set of size, shape, morphology and normalized color descriptors is computed. The final classification is implemented using a k-nearest neighbors classifier that yields a correct diagnosis in 88% of the cases. Later, Schmid-Saugeon, Guillodb, and Thirana (2003) also presented an integrated system aiming at the differential diagnosis of melanoma against melanocytic naevi. Their system removes undesirable artifacts from the images (i.e. hair), segments the lesions from the healthy skin and finally computes a set of features conforming to the Asymmetry criterion of the ABCD diagnostic rule. This system achieved very good segmentation results, yet inferior classification results even compared to earlier attempts, reaching 73.2% total diagnostic accuracy. Burroni et al. (2005) conducted a study focusing on the differentiation of in situ melanoma lesion from dysplastic naevi. They used 174 lesion images and different features regarding color, texture and the geometry of lesions and obtained a classification rate of 71.8%. The ADAM system (Stanganelli et al., 2005) achieved slightly better results reaching 80% for both sensitivity and specificity using Asymmetry and

* Corresponding author.
E-mail addresses: i.e.giotis@rug.nl (I. Giotis), n.molders@umcg.nl (N. Molders), sander.land@wolfson.ox.ac.uk (S. Land), m.biehl@rug.nl (M. Biehl), m.f.jonkman@umcg.nl (M.F. Jonkman), n.petkov@rug.nl (N. Petkov).

http://dx.doi.org/10.1016/j.eswa.2015.04.034
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

Boundary descriptors and a Support Vector Machine (SVM) classifier. The study of Iyatomi et al. (2008) is also using dermoscopic images and reports 100% specificity and 95.9% sensitivity. One of the latest systems presented in the literatures is that of Abbas, Emre Celebi, Garcia, and Ahmad (2013) that achieves 88.2% sensitivity and 91.3% specificity on a dataset of 120 dermoscopic images.

Lately, a new trend in the computer-assisted diagnosis of early melanoma has emerged in the form of web and mobile phone applications. These applications do not use dermoscopic images as input but rather images taken with standard digital cameras. One of the first and most successful applications in this area, according to user and press reviews, is SpotMole (Munteanu & Cooclea, 2009). The imaging and pattern recognition techniques that comprise the analysis steps of SpotMole are also conforming with the ABCD diagnostic set of rules. A newer attempt that also received positive reviews is called MelApp (Health Discovery Corporation, 2011). Prior to its commercial release MelApp has been assessed with a set of images from the Johns Hopkins University Medical Center. MelApp does not provide a classification result but rather gives an indication of (low, medium or high) risk of melanoma, thus rendering the evaluation of its effectiveness in terms of correct classification difficult.

Korotkov and Garcia (2012) recently presented an extensive and comprehensive review regarding the computerized analysis of pigmented skin lesions. Their work classifies relevant studies according to the type of images used (dermoscopic or clinical), the amount of lesions present in an image, the algorithms used for processing etc. In this study they reach the important conclusion that a publicly available, benchmark dataset is essential in order for such systems to gain wider acceptance among patients and clinicians. Among the works focusing on clinical images the one of Zagrouba and Barhoumi (2011) presents a system with similar properties as MED-NODE; a noise removal preprocessing step, localization of the regions of interest for feature extraction etc. This system focuses on the extraction of features that will adequately quantify the ABCD diagnostic rule, i.e. compactness and asymmetry indices, color homogeneity, etc. All those features are subsequently used to train a mono-layer perceptron classifier and yield a system with sensitivity of 75.1% and specificity of 83.1% on a dataset consisting of 125 benign and 75 malignant lesions. The system proposed by Alcon et al. (2009) also uses simple digital images as well as context knowledge, such as skin type, age, gender, and the affected body part in order to classify lesions as benign or malignant. This system also attempts to extract features conforming with the ABCD rule and includes steps for preprocessing and segmenting the lesions. The extracted features are used in a logistic regression classifier with a processing step for correlation-based feature selection and yielded a sensitivity of 94%, and specificity of 68% on a dataset of 45 benign and 107 malignant lesions.

In this contribution we present a novel computer-assisted diagnostic (expert) system for the differentiation of melanoma from nevocellular naevi using non-dermoscopic images, that we call MED-NODE (computer-assisted MElanoma Diagnosis from NOn-DErmoscopic images). The vital contributions of MED-NODE lie mainly in the use of simple digital images (instead of dermoscopic images) which are much easier to obtain and in the integration of the symptoms observed by the examining physician in the lesion classification process. We present a complete framework for the accurate detection of melanoma lesions that incorporates three vital properties: (a) it is able to effectively handle noise and illumination effects, thus simplifying the image acquisition process, (b) it is able to automatically locate informative regions within a lesion and extract the corresponding color-texture features, thus minimizing the user input required and (c) it incorporates medical domain knowledge in the form of annotations with regard to the presence of simple visual attributes of the lesions. Such a system can be significantly beneficial to the first line of health care (general practitioners) where the dermatological equipment is not available, the dermatology specialists that need to extend their case knowledge base, but also to the patients themselves. This paper is organized as follows: Section 2 describes in detail the materials and methods used in this study, Section 3 presents the experimental results, Section 4 discusses open issues and prospects for future work and finally Section 5 presents the conclusions drawn.

2. Materials and methods

The MED-NODE system relies on three different sources of information regarding each patient case: lesion color, lesion texture and visual diagnostic attributes. These sources are initially handled independently and in the final stage of decision making they contribute equally to the classification of a lesion. The equal contribution of all three data sources is chosen here both due to the lack of prior knowledge with regard to the importance of the feature sets and its simplicity and intuitive clarity. Color and color texture characteristics for each lesion are automatically extracted from the digital images of patient cases, whereas the presence or absence of a set of visual attributes widely used in dermatological diagnosis is determined by the examining physician. The final classification decision for a test case is taken based on a majority vote among the predictions stemming from the three sources.

In the literature there exist different ways of combining cooperating classifiers with inconclusive results as to whether one method is performing better than the rest. Kittler, Hatef, Duin, and Matas (1998) investigated the product, the sum, the maximum, the minimum, the median, and the majority voting methods for general pattern recognition tasks and found that the sum method outperformed other combination methods. However, Suárez-Cuenca, Guo, and Li (2011) concluded that for their computer-aided diagnosis system the majority-vote rule achieved higher performance levels than other combination methods. In this contribution we chose the majority vote rule mainly due to its intuitive simplicity and probity.

2.1. Dataset

Our dataset consists of 70 melanoma and 100 naevus images from the digital image archive of the Department of Dermatology of the University Medical Center Groningen (UMCG). The photos of the lesions were taken with a Nikon D3 or Nikon D1x body and a Nikkor 2.8/105 mm micro lens. In most cases (95%) the distance between the lens and the lesion is approximately 33 cm. The lighting conditions were set using two Multiblitz Variolite 600 flash units with a color temperature equal to 5200 Kelvin. All images are part of the much larger dermatology digital archive that consists of more than 50,000 images of different types of lesions. For the development of the MED-NODE expert system we chose for a smaller dataset of randomly chosen images where the underlying patient cases are completely unidentifiable. This dataset contains only superficial spreading melanoma and naevi, thus avoiding seborroic warts, spitz naevi, and acrolentigineus/nodular melanomas. The images of pigmented skin lesions originate only from patients of caucasian origin, who constitute the vast majority of the population in the Netherlands. For each picture the available diagnosis has been verified by the medical correspondence of the Department of Dermatology. In order to further ensure the soundness of the data, the following selecting criteria have been employed:

1. Every picture must originate from a different patient (apart from cases where a disease looks clearly different at different parts of the body, which can be included in the dataset).
2. Each picture must be sharp and properly exposed, so it can be appropriately annotated.
3. Each picture must be representative of the group it belongs to. Rare clinical variants, already treated and/or secondarily infected skin diseases are not included in the dataset.

From each image a region of interest is manually selected, that contains both healthy skin and (part of) a lesion, but without any distracting elements (clothes, background, jewelry, etc). The scale and size of the selected regions differs per image and are dependent on the size of the original image, which in turn depends on the size of the lesion and the location of the lesion on the body. Hair is considered a distracting element as well and has been manually removed using the Dullrazor software (Lee, Ng, Gallagher, Coldman, & McLean, 1997). Considering the explicit need for a public, benchmark dataset in the field (Alcon et al., 2009), we made publicly available the anonymized, preprocessed images at http://www.cs.rug.nl/~imaging/databases/melanoma_naevi.

2.2. Lesion segmentation

The selected regions of interest are segmented in healthy and lesioned skin areas using the k-means clustering algorithm, with $k = 2$. The $k = 2$ clusters represent the healthy and lesioned skin respectively. This way every pixel within the region of interest is classified as part of the lesion or healthy skin. However, a series of pre-processing steps are necessary before the clustering is applied in order for it to effectively handle noise and illumination effects (Land, 2009). The images in our dataset are often noisy and have non-uniform lighting. Therefore, the results of traditional segmentation approaches would be dominated by lighting effects. To achieve proper segmentation, an algorithm was devised to remove lighting effects' gradients from images which are affected by non-uniform lighting, while leaving other images unaffected.

The segmentation is implemented in the HSV color space, where such effects of illumination usually present themselves as a clear gradient in the saturation and value channels, which makes it easier to detect and eliminate them. HSV stands for hue (H), saturation (S), and value (V), with value being an equivalent for greyscale intensity. The HSV colorspace was chosen since it is based upon the way colors are perceived in human vision. For further discussion of appropriate color spaces we refer to Bosman, Petkov, and Jonkman (2010). At first, Gaussian smoothing with $\sigma = 5$ is applied to the original image to remove noise. Subsequently, the squared gradient magnitude of an image is determined as the sum of the squared gradient magnitude of all color channels. Next, a mask is constructed by grouping all pixels with a squared gradient magnitude below the median. The gradients of pixels in this mask are considered the result of illumination effects rather than real edges. Excluding all pixels in this mask, the mean gradients in the saturation and value channels are determined. Finally, we approximate the background of the image in these two channels, as an image with precisely this mean gradient and a pixel value equal to 0 in the center. The estimated background is subtracted from the saturation and value channels in the original non-smoothed image. The corrected saturation and value channels are then combined with the original hue channel. Additional noisy features are removed using a Kuwahara smoothing filter (Kuwahara, Hachimura, Eiho, & Kinoshita, 1976) that is able to preserve edges.

Finally, k-means clustering is applied in the HSV color space segmenting each image in healthy and lesioned sections. Since the hue channel represents an angle, values close to the minimum and maximum should be considered close together, which does not happen when the Euclidean distance metric is used. Therefore, the distance on the hue channel is defined as:

$$d_h(h_a - h_b) = min(|h_a - h_b|, 1 - |h_a - h_b|) \tag{1}$$

and the total distance between two colors in the HSV color space as:

$$d_{hsv} = ||c * d_h, s_a - s_b, v_a - v_b|| \tag{2}$$

The problem of defining the mean of several angles (for the hue channel) is solved by converting each angle ω to the unit circle $\omega \rightarrow e^{i\omega}$, computing the mean of all unit circles, and using the argument of this mean as the mean value of the angles. Fig. 1 presents two examples of images from our dataset (left) and the respective segmentation results (right).

2.3. Classification using color descriptors

With regard to color, a simple set of statistical features in the RGB color space is used. The mean color of the healthy skin region, as extracted by the segmentation process, is computed and then subtracted from each pixel of the lesioned skin region. Subsequently, the mean and standard deviation of all color channels of the original and the normalized lesioned region are computed comprising a 12-dimensional feature vector. Finally, the Cluster-based Adaptive Metric (CLAM) classifier (Giotis & Petkov, 2012) is utilized to assign each picture to either melanoma or nevocellular naevus.

CLAM is a prototype-based classifier which effectively addresses common drawbacks of similar classification methods, such as the tuning of input parameters, without requiring a computationally intensive training phase. The motivation for using CLAM in this system stems mainly from these advantages with regard to speed and ease-of-use. The algorithm comprises three main steps. First, the number of underlying sub-groups (clusters) within each class of the training set is estimated using the gap statistic test (Tibshirani, Walther, & Hastie, 2001). This test compares the within-cluster dispersion of a dataset comprising k clusters with the dispersion of a reference set R_i that comprises only one cluster. The optimal number of clusters $\hat{k}$, is then the value of k for which the logarithm of the dispersion index of the real

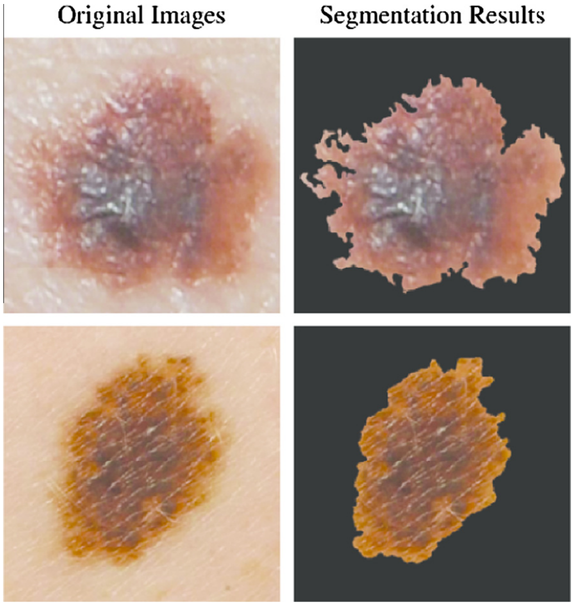

Fig. 1. Segmentation results.

dataset falls the furthest below the respective one of the reference set. Let C_{ij} be the jth cluster of the ith class. In the second step, a mean μ_{ij} and a covariance matrix $\hat{\Sigma}_{ij}$ for each cluster C_{ij} are estimated and used to compute the parameters of a distance metric d_{ij} with quadratic form to that cluster. The last step of the CLAM algorithm computes the probability of cluster-membership for a testing data point x and cluster C_{ij} as a Bayesian posterior probability, taking into account the distances to the cluster and the number of data points in it as shown in Eq. (3):

$$P(C_{ij}|x) \propto \frac{1}{\sqrt{\hat{\Sigma}_{ij}}} e^{-\frac{1}{2}d_{ij}(x)} \tag{3}$$

where

$$d_{ij}(x) = (x - \mu_{ij})\hat{\Sigma}_{ij}(x - \mu_{ij})^T \tag{4}$$

is the adapted quadratic form distance computed in the second step. The class-membership probability of the data point x for the ith class is computed as the sum of the relevant cluster-membership probabilities.

2.4. Classification using color-texture descriptors

The texture features are extracted from the RGB colorspace representation of the images using the Color Image Analysis Learning Vector Quantization (CIA-LVQ) (Giotis, Bunte, Petkov, & Biehl, 2013a) supervised learning framework, which is modified accordingly in order to decrease various sorts of bias. We refer to the modified variant as Unbiased CIA-LVQ (UCIA-LVQ). It is important to note that with regard to color-texture descriptors the whole image is being used and not only the previously segmented lesion regions.

CIA-LVQ is a color texture classification and recognition algorithm based on the Generalized Matrices Learning Vector Quantization (GMLVQ) (Schneider, Biehl, & Hammer, 2009). Learning Vector Quantization (LVQ) is a supervised prototype-based classification method (Kohonen, Schroeder, & Huang, 2001). The training is based on data points $x \in \Re^D$ and their corresponding label information $y^k \in 1, \ldots, C$, where D denotes the dimension of the feature vectors and C the number of classes. A set of prototypes is characterized by their location in the feature space $w^i \in \Re^D$ and the respective class label $c(w^i) \in 1, \ldots, C$. Classification is implemented as a winner-takes-all scheme. For this purpose, a possibly parameterized dissimilarity measure d^Ω is defined, where Ω specifies the metric parameters which can be adapted during training. Given d^Ω(x,w), any data point x is assigned to the class label $c(w^i)$ of the closest prototype w^i with $d^\Omega(x, w^i) < d^\Omega(x, w^j)$ for all $i \neq j$.

In the case of CIA-LVQ let **D** be a dataset of color images of a priorly known size $p \times p$ that belong to C different classes and a bank of filter kernels **G**. The algorithm learns one or more matrices Ω_k that transform the color images into single-channel, intensity images, a set of optimized kernels $\widehat{G}_k$ and a set of prototypes w^k such that the filter responses of the transformed images will yield the best possible classification. The most important modification of the standard CIA-LVQ relates to the initialization of the filter bank used. The original algorithm relies on kernels initialized as Gabor filters in order to identify textural patterns. In the case of skin lesions, however, the surface descriptors might not relate closely to the strict notion of texture as used in computer vision. Therefore, we initialize the sum of all kernels G as a random 2-dimensional matrix of size $p \times p$, smoothed by a 2D-Gaussian function with $\sigma = 1$. Thereafter, G is normalized so that the sum of its elements in the spatial domain is equal to zero in order to ensure the existence of equal excitatory and inhibitory regions in the filter kernel. In this fashion the adapted kernels will still be able to identify contrast changes as surface characteristics, without being biased from the Gaussian form of the Gabor filters. The second modification is related to the image patches (blobs) we use to formulate the training and evaluation sets for UCIA-LVQ. In order to avoid drawing patches from non-informative regions (i.e. homogeneous regions) we employ a keypoint detector as a preprocessing step. In this contribution we opt for the salient point detector from Walther and Koch (2006), which is an extension of the visual attention model of Itti, Koch, and Niebur (1998). Finally, the updates of the parameters in the UCIA-LVQ variant are implemented in batch mode. This means that the prototypes, transformations matrices and filter kernels are not adapted after every data point is presented but rather once at the end of every iteration using normalized gradients (Papari, Bunte, & Biehl, 2011).

Every image in our dataset is represented by 50 patches of size 15×15 pixels. The 50 most interesting positions (pixels) in every image are computed using the salient point detector and subsequently 15×15 patches are drawn centered around those positions. This results in possibly overlapping patches on highly informative parts of the image and is likely to leave completely unsampled parts of the image that are not considered informative, i.e. large regions of homogeneously colored healthy skin. We represent every class with one prototype and train the localized version of the UCIA-LVQ classifier for 500 epochs (iterations). Subsequently, in order to assign an image to a disease, we take into account the classification of all 50 patches drawn from this image. We first rank the patches in ascending order of their distances to the closest prototype and then apply a weighted majority vote with weights inversely proportional to this ranking. In this fashion, the influence in the majority vote of the patch that ranks first (i.e. the patch that lies the closest to a prototype among all 50 patches of an image) will be significantly higher compared to the influence of the patch that ranks last. The exact size of the image patches is a parameter that can be set by the user. It is, however, related to the size of the input images and the dimensionality of the vector space used to train the LVQ module. In our case, 15×15 patches ensure that this dimensionality will not be inconsiderately large given the number of images at our disposal and that a reasonable amount of patches can be drawn from every input image for the final step of majority voting to be meaningful.

2.5. Classification using diagnostic visual attributes

In this contribution we use physician annotations for skin lesions according to the condensed lexicon by Giotis, Visser, Jonkman, and Petkov (2013b) which is based on the PROVOKE (Sillevis Smit, van Everdingen, Starink, & van der Horst, 2009) system, used for the training of dermatologists in the Netherlands. The visual attributes used in the PROVOKE system are organized in 10 generic groups that we call aspects as shown in Table 1. An attribute that concerns a given aspect can be assigned to a case. Each attribute is a binary feature in our terminology, assigned the value 1 when it is present and the value 0 when it is not present on a given lesion. The condensed annotation lexicon is developed using an automated way to answer the question of feature discriminability for dermatological diagnosis using an information theoretical approach. Our approach to the estimation of the discriminative power of the attributes of the PROVOKE lexicon is based on the probability $p(f|\omega)$ that a certain disease ω, will cause the presence of a given attribute f. Subsequently, we use the entropy of the distribution of the probabilities of occurrence of f across different diseases to measure the attributes discriminative power. A uniform probability (high entropy) means that a given attribute is equally

Table 1
Condensed dermatological lexicon.

Aspect	Attributes
1. Part of the Body	**Head**, Neck, Trunk front, Trunk back, Arm, Hand, Buttocks, Leg, **Foot**
2. Spatial Arrangement	Corymbiform, **Annular**, Linear, Herpetiform, Disseminated, Diffuse, Discrete, Reticular, **Confluent**, Follicular, **Circinate**, **Concentric**, Target shape, **Solitary**
3. Number	**One**, Few (<5), Several (<10), Many (>10)
4. Size	Extra small (1–3 mm), Small (3–10 mm), Medium (1–3 cm), **Large (3–5 cm)**, Extra Large (>5 cm)
5. 2-D Shape	Round, Oval, Polygonal, Polycyclic, Rectangular, Linear, Gyrated, Dendritic, Irregular, Annular, Arciform
6. 3-D Shape	Spherical, Spherical with indentation, Hemispheric, Flat, Tapered, Blunt, **Not Elevated**, Rough, **Raised Edge**, Pendiculate
7. Boundary Sharpness	Sharp, Diffuse
8. Color	Normal (same color as healthy skin), White, **Red**, Blue, **Brown**, Black, Gray, **Multi-color**
9. Morphological Group	Erythema, **Dyschromia**, **Papular**, Urticarial, Nodular, Tumor, **Erythema-papulo-squamous (Dermatosis)**, Pustular (Dermatosis), Vesicular/Bullous, **Ulcerative**
10. Surface	**Smooth**, **Coarse**, Folded, Wrinkled, **Verrucous**, **Papillomatous**, Moist, Purulent

likely to be observed in all the considered diseases, whereas low entropy distributions characterize highly discriminative features. The 22 attributes achieving the highest discriminability scores, shown in bold letters in Table 1, comprise the final condensed lexicon of physician annotations. The annotations for our dataset were provided by a dermatology specialist and a medical school graduate under the supervision of the head of the Dept. of Dermatology of the University Medical Centre Groningen. We use these attributes to form binary feature vectors based on the presence or not of an attribute in a skin lesion and use these features as input for a naive Bayesian classifier.

3. Experimental results

We test the performance of the proposed system on the sample data from the University Medical Center Groningen. We randomly split the images into training and evaluation sets with an approximate 25%/75% ratio. This leads to a total of 45 training (20 for melanoma and 25 for nevocellular naevus) and 125 evaluation (50 for melanoma and 75 for nevocellular naevus) feature vectors for the color and the annotated features. As to the color-texture features we obtain 2250 training feature vectors and 6250 evaluation vectors. Fig. 2 depicts a few representative examples of lesion images that comprise our dataset.

We assign lesions to diseases based on their color using the CLAM classifier. The accuracy obtained in the evaluation set is 73.6%. More specifically CLAM is able to correctly classify 37 out of 50 cases of melanoma (74%) and 54 out of 75 (72%) cases of nevocellular naevus. With regard to color-texture, every 15 × 15 patch is initially assigned to one of the two studied conditions using a distance parameterized with the learned parameters w^k, Ω_k and $\widehat{G}_k$. Fig. 3 presents the final form of the prototypical patches for each skin condition. Using this method we obtain

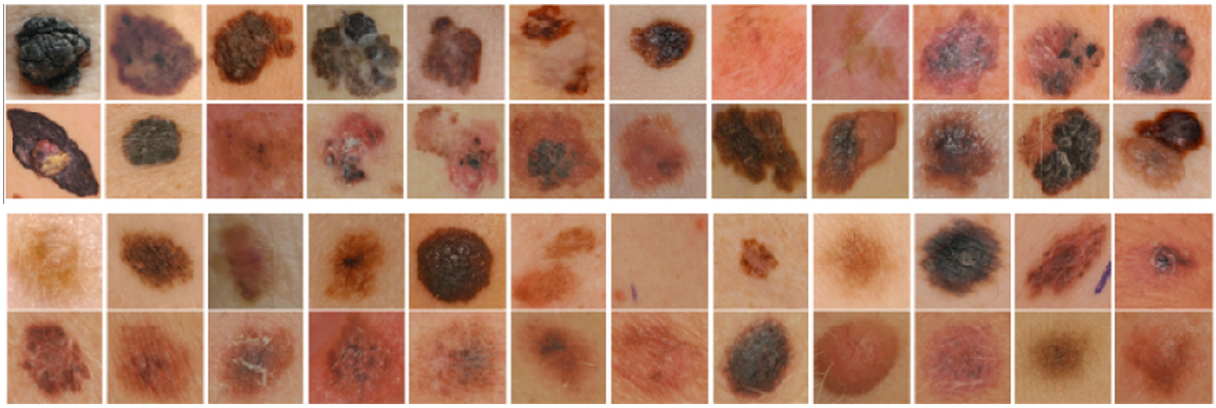

Fig. 2. Example images from the UMCG dataset: the first two rows depict 24 melanoma cases, whereas the third and fourth rows contain an equal number of benign naevus cases.

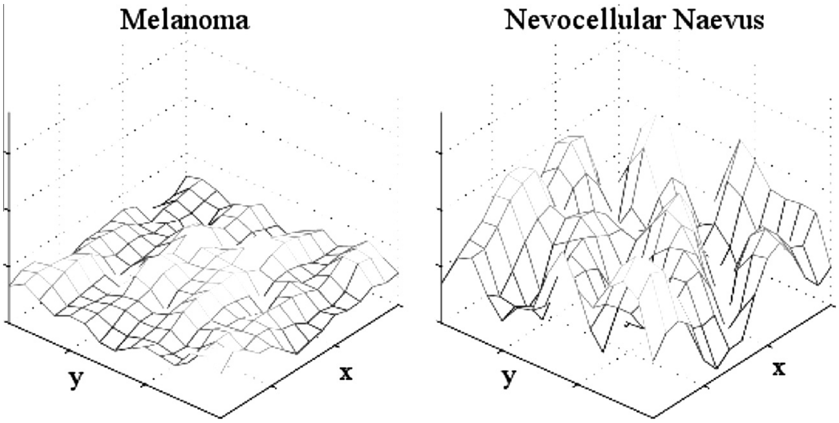

Fig. 3. Magnitude of a prototypical lesion patch of melanoma (left) and nevocellular naevus (right).

Table 2
Diagnostic Accuracy (sensitivity–specificity) Results.

		Descriptor used			
		Color	Texture	Annotations	Full system
Skin condition	Melanoma (sensitivity)	0.74	0.62	0.78	0.80
	Nevocellular naevus (specificity)	0.72	0.85	0.59	0.81
	Total accuracy	0.73	0.76	0.66	0.81

accuracy 62% for melanoma and 85% for nevocellular naevus. Finally, using the annotations of the 22 visual attributes presented in Table 1 and the Naive Bayes classifier we are able to obtain 78% diagnosis accuracy for melanoma cases and 59% for nevocellular naevi.

At this stage the output of our method consists of three diagnosis predictions for each image in the evaluation set. The final assignment of these images to one of the two diseases is implemented with a majority vote of the three predictions. Table 2 summarizes the diagnostic accuracy results. For both studied skin conditions the complete MED-NODE system is able to reach class-wise accuracies higher than 80% for previously unseen images. Note, that class-wise accuracy rates are actually equivalent to sensitivity (melanoma) and specificity (naevus) rates. The detailed results also demonstrate why the complete system is more powerful than the individual descriptors. Color and annotated features are able to reach the highest result with regard to melanoma individually. However, they both fail to identify benign naevi to the same level. Conversely, the color-texture features perform better in relation to the class of nevocellular naevi. Consequently, only the combination of descriptors is able to reach the highest overall classification rate.

We estimate the interval for our estimations at the 95% confidence level using the central limit theorem. The upper and lower bounds of the estimated classwise accuracies are given by:

$$acc_i = \pm 1.96\sqrt{\frac{1}{n} * acc_i * (1 - acc_i)} \tag{5}$$

In this fashion the sensitivity of the MED-NODE system falls with 95% confidence within the [0.66, 0.90] interval and its specificity within the [0.71, 0.89] interval.

We compare the results of MED-NODE with those of the web application version of SpotMole using the same evaluation set. We used the web-based version of SpotMole (SpotMoleJ), which is a remotely hosted java applet and thus platform-independent. SpotMole achieves a total diagnostic accuracy of 67.2%, performing similarly to MED-NODE with regard to melanoma lesions (82% accuracy), whereas regarding nevocellular naevi it raises too many false positives, correctly recognizing only 57.3% of the lesions. Similarly, the estimated sensitivity for SpotMole falls with 95% confidence within the [0.67, 0.91] interval and its specificity within the [0.45, 0.69] interval. Although the exact underlying features used to classify lesions are not known, SpotMole seems to rely heavily on the automatically computed border of the lesions, which in various cases in our evaluation set was not accurate. The irregularities of the automatically computed lesion borders can be an important factor that leads to false positives and thus low specificity, given that melanoma lesions are known to be asymmetrical and with highly irregular borders.

We also compare the results of MED-NODE with the work of Zagrouba and Barhoumi (2011) using the implementation available at (Rosa, 2014) and the exact same training and evaluation sets. This system achieves a total diagnostic accuracy of 70.4%, performing worse than the other two with regard to melanoma lesions (46% accuracy), whereas regarding nevocellular naevi it has the best performance, correctly recognizing 86.7% of the lesions. Hence, the estimated sensitivity falls with 95% confidence within the [0.32, 0.61] interval and its specificity within the [0.77, 0.93] interval. In order to further clarify the advantages of the proposed approach, we compare all three systems in terms of positive (PPV) and negative (NPV) predictive values. The results with regard to MED-NODE are presented in Table 3. MED-NODE has a PPV of 74.1% and NPV equal to 85.9%, raising 14 false alarms for melanoma and disregarding as naevi 10 melanoma cases in the evaluation set of 125 images, whereas SpotMole raises 32 false alarms, reaching only 56.1% PPV and disregards as naevi 9 melanoma lesions reaching NPV equal to 82.7%. The system of Zagrouba and Barhoumi (2011) also performs worse than MED-NODE with PPV equal to 69.7% and NPV equal to 70.7%, hence raising 10 false alarms but also disregarding 27 melanomas as benign lesion.

Finally, we present for completeness and in order to facilitate for the reader an overview of the systems shortcomings, a few examples of misclassified images from both classes in Fig. 4.

4. Discussion

The MED-NODE system shows the potential to outperform state-of-the-art methods for computer-assisted melanoma diagnosis, using solely non-dermoscopic images. The methods of

Table 3
MED-NODE positive and negative predictive values.

	Descriptor used			
	Color	Texture	Annotations	Full system
PPV	0.638	0.738	0.557	0.741
NPV	0.806	0.771	0.800	0.859

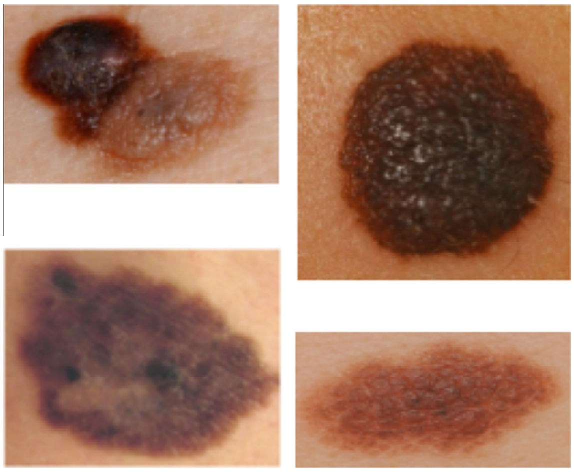

Fig. 4. Examples of melanoma (left) and naevus (right) lesions wrongly classified by the MED-NODE system.

Schmid-Saugeon et al. (2003) and Burroni et al. (2005) reached lower diagnostic accuracy, 73.2% and 71.8% respectively, and only the works of Ganster et al. (2001), Iyatomi et al. (2008) and Abbas et al. (2013) report higher results than MED-NODE. However, one should note that these quantitative results are rendered using different datasets and that the images used in those studies are dermoscopic and not clinical digital images. Due to this fundamental difference in the scope of those systems a straightforward comparison with any of these methods is not feasible or meaningful in the case of MED-NODE. Non-dermoscopic images are much easier to obtain but nevertheless subject to more noise and illumination effects. Therefore, our dataset would not be a suitable input for techniques designed on the principle that lesion images are unimpeded by skin surface reflections and other types of noise.

The system of Zagrouba and Barhoumi (2011) performs on our dataset (70.4%) comparably to what the authors mention in their contribution (79.1%) and suffers mainly from too many false negatives; that is melanoma lesions that are classified as benign. The features used in this system are derived from the well-known ABCD diagnostic rule of dermatology (Asymmetry, Border irregularity, Color, Diameter) attempting to model with computer vision techniques the diagnostic process of trained physicians.

The choice of a commercial application, such as SpotMole, as a baseline for comparison, is consciously made given the public appeal such applications have recently gained. Applications such as SpotMole, DoctorMole (Revosoft, 2014), etc. are publicly available and easy to use, and they are becoming more and more popular among the general public ($\sim$100,000 users only for Android devices), yet they fail to provide a solid scientific background as to why their classification results are reliable. Furthermore, SpotMole claims to comprise of imaging and pattern recognition techniques are also conforming with the ABCD diagnostic set of rules.

It is important to observe that both techniques we compare MED-NODE with, use features in accordance with the ABCD diagnostic rule. The resulting inconsistencies in their classification performance can be partly due to this fact. Both systems seem to depend very strongly on the computation of the lesion border which in many cases is not very accurate, due to errors at the lesion segmentation step. Furthermore, both systems attempt to extract one set of features from the whole lesion, which can also prove to be an erroneous approach. On the contrary MED-NODE does not attempt to model dermatology domain knowledge using computer vision techniques. The proposed system focuses on (a) using human knowledge through a set of simple, comprehensive and discriminative features and (b) exploiting properties (color and texture) where pattern recognition techniques have already been proven effective and that can be applied in this field. The superior results of MED-NODE, regardless of the very small set of images used to train the system support empirically this argument. Finally, our contribution combines the ease-of-use and the absence of need for expensive equipment while providing exactly the consistent methodology and mathematical foundation that ensure the quality of its results.

An issue that needs to be addressed at this point concerns the prior probabilities of a skin mole being melanoma or benign naevus. The prior probability of a mole being a benign naevus is much higher and therefore such a system would yield a large amount of false positives if used to screen random skin moles. Its use becomes much more sensible in cases where there is some indication that a mole is suspicious, e.g. it appeared recently or grew too rapidly (evolution), so that the prior probabilities of it being a benign mole or a melanoma are equal or at least comparable. The MED-NODE system can be further enhanced to use automatically extracted descriptors for the spatial arrangement and morphology of lesions. These aspects are also proven important in the diagnostic process and in this fashion the manual input required could be further reduced. Similar systems, using the principle of different descriptor modalities, for other groups of diseases that are often mistaken for one another (i.e. types of psoriasis and keratosis) could be implemented based on the information sources used in this work and become part of a more generic computer-assisted diagnosis tool for dermatology. Finally, in order to reduce bias and improve the diagnostic accuracy of MED-NODE more advanced techniques than majority vote could be tested for the combination of classifiers, such as boosting (Freund & Schapire, 1997) or logistic regression (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996).

5. Conclusion

In this contribution we present a novel, expert system for computer-assisted diagnosis of skin cancer, called MED-NODE. It aims to differentiate melanoma from nevocellular naevi using simple digital images of lesions and three different types of descriptors. Descriptors regarding the color and texture of lesions are automatically extracted, and together with a set of manual annotations related to additional visual attributes of the lesions, are used to predict whether a lesion is melanoma. The proposed system performs well, achieving high diagnostic accuracy results (81%) despite the small size of our dataset. MED-NODE consistently outperforms the SpotMole system and the system proposed by Zagrouba and Barhoumi (2011) on the same set of images and does not show bias towards any of the two classes.

With respect to the area of computer-aided diagnosis systems for dermatology the contribution of our work is twofold: (a) we present a system that is able to successfully differentiate melanoma from common nevi using simple, macroscopic digital images instead of dermoscopy and (b) we present the framework to include physician-observed attributes of a lesion in the automated recognition process. The simple requirements regarding the input images together with the completeness and the soundness of the features used to describe the lesion characteristics and the short processing time needed, allow for this system to be utilized in different ways: as a tool to facilitate initial melanoma screenings, as an interface to conduct dermatological consultations remotely (teledermatology) (Ebner et al., 2006; Perednia & Brown, 1995) or even as a means for the training of dermatology specialists. Finally, we provide a benchmark, publicly available set of images with verified diagnosis, in order to facilitate further research and assist in the wider recognition of systems akin to MED-NODE. In the general area of expert systems this contribution clearly demonstrates the added value of state-of-the-art machine learning algorithms and the combination thereof in a discipline of healthcare, a field of significant public interest. Furthermore, the techniques used in the classification process are proven able to tackle the common issue of noisy data (a) by learning descriptive prototypical representatives of the two classes (melanoma and nevi) and (b) by utilizing suitable localization techniques to extract features (segmentation, keypoint detection).

The main limitation of this study lies in the size of the dataset used to evaluate its performance. Although datasets of similar size are not uncommon in this area (Abbas et al., 2013; Iyatomi et al., 2008), due to the sensitive nature of the data and the difficulty to obtain labeled images from the medical community (time consuming task, uncertainty in diagnosis, etc.), such a system should be further evaluated using a larger set of data before being tested in a clinical environment.

Future research shall focus on the expansion of the proposed system in terms of (a) different types of visual characteristics such as shape, spatial arrangement etc. and (b) other challenging dermatological diagnosis tasks such as the differentiation of psoriasis and eczema, diseases that also present visual similarities. The inclusion of more than two diseases in the classification process of the system is another interesting possibility in the same direction. Finally, in a broader context the proposed system could be expanded to the 3D imaging domain, which could render it beneficial in other medical disciplines as well.

Acknowledgments

The authors would like to thank Margaretha Visser for assisting during the selection and annotation of the images comprising the dataset used in this study and Dr. Kerstin Bunte for the fruitful discussions during the development of the UCIA-LVQ algorithm.

References

Abbas, Q., Emre Celebi, M., Garcia, I. F., & Ahmad, W. (2013). Melanoma recognition framework based on expert definition of abcd for dermoscopic images. *Skin Research and Technology, 19*, e93–e102. http://dx.doi.org/10.1111/j.1600-0846.2012.00614.x.

Alcon, J., Ciuhu, C., ten Kate, W., Heinrich, A., Uzunbajakava, N., Krekels, G., et al. (2009). Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis. *IEEE Journal of Selected Topics in Signal Processing, 3*, 14–25. http://dx.doi.org/10.1109/JSTSP.2008.2011156.

Bosman, H. H., Petkov, N., & Jonkman, M. F. (2010). Comparison of color representations for content-based image retrieval in dermatology. *Skin Research and Technology, 16*, 109–113.

Burroni, M., Sbano, P., Cevenini, G., Risulo, M., Dell'eva, G., Barbini, P., et al. (2005). Dysplastic naevus vs.in situ melanoma: Digital dermoscopy analysis. *British Journal of Dermatology, 152*, 679–684.

Ebner, C., Gabler, G., Massone, C., Hofmann-Wellenhof, R., Lozzi, G. P., Wurm, E., et al. (2006). Mobile teledermatology coming of age. *e & i Elektrotechnik und Informationstechnik, 123*, 148–151. http://dx.doi.org/10.1007/s00502-006-0333.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139. http://dx.doi.org/10.1006/jcss.1997.1504.

Friedman, R. J., Rigel, D. S., & Kopf, A. W. (1985). Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians, 35*, 130–151. http://dx.doi.org/10.3322/canjclin.35.3.130.

Ganster, H., Pinz, A., Rohrer, R., Wildling, E., Binder, M., & Kittler, H. (2001). Automated melanoma recognition. *IEEE Transactions on Medical Imaging, 20*, 233–239.

Giotis, I., Bunte, K., Petkov, N., & Biehl, M. (2013a). Adaptive matrices and filters for color texture classification. *Journal of Mathematical Imaging and Vision, 47*, 79–92. http://dx.doi.org/10.1007/s10851-012-0356-9.

Giotis, I., & Petkov, N. (2012). Cluster-based adaptive metric classification. *Neurocomputing, 81*, 33–40. http://dx.doi.org/10.1016/j.neucom.2011.10.018.

Giotis, I., Visser, M., Jonkman, M., & Petkov, N. (2013b). Discriminative power of visual attributes in dermatology. *Skin Research and Technology, 19*, e123–e131.

Health Discovery Corporation. (2011). Worlds first svm-based image analysis iphone app for melanoma risk assessment, melapp, launched by health discovery corporation. <http://www.healthdiscoverycorp.com/pr/july06_11.html>.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259. http://dx.doi.org/10.1109/34.730558.

Iyatomi, H., Oka, H., Celebi, M. E., Ogawa, K., Argenziano, G., Soyer, H. P., et al. (2008). Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin. *Journal of Investigative Dermatology, 128*, 2049–2054.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 226–239. http://dx.doi.org/10.1109/34.667881.

Kohonen, T., Schroeder, M. R., & Huang, T. S. (Eds.). (2001). *Self-organizing maps* (3rd ed.. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Korotkov, K., & Garcia, R. (2012). Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine, 56*, 69–90.

Kuwahara, M., Hachimura, K., Eiho, S., & Kinoshita, M. (1976). Processing of ri-angiocardiographic images. In K. Preston, Preston & M. Onoe (Eds.), *Digital processing of biomedical images* (pp. 187–202). US: Springer. http://dx.doi.org/10.1007/978-1-4684-0769-3_13.

Land, S. (2009). Content-based image retrieval in dermatology (Master's thesis). Rijksuniversiteit Groningen the Netherlands.

Lee, T., Ng, V., Gallagher, R., Coldman, A., & McLean, D. (1997). DullRazor: A software approach to hair removal from images. *Computers in Biology and Medicine, 27*, 533–543.

Marks, R. (2000). Epidemiology of melanoma. *Clinical and Experimental Dermatology, 25*, 459–463. http://dx.doi.org/10.1046/j.1365-2230.2000.00693.x.

Munteanu, C., & Cooclea, S. (2009). Spotmole – melanoma control system. <http://www.spotmole.com/>.

Papari, G., Bunte, K., & Biehl, M. (2011). Waypoint averaging and step size control in learning by gradient descent. Technical Report MLR-2011-06 Leipzig University.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*, 1373–1379.

Perednia, D. A., & Brown, N. A. (1995). Teledermatology: One application of telemedicine. *Bulletin of the Medical Library Association, 83*, 42–47.

Revosoft. (2014). Doctor mole – app to check skin cancer by dr mole. <http://www.doctormole.com/>.

Rosa, L. (2014). Advanced source code. com – melanoma recognition system. <http://www.advancedsourcecode.com/melanomarec.asp>.

Rosado, B., Menzies, S., Harbauer, A., Pehamberger, H., Wolff, K., Binder, M., et al. (2003). Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis. *Archives of Dermatology, 139*, 361–367.

Schmid-Saugeon, P., Guillodb, J., & Thirana, J. P. (2003). Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics, 27*, 65–78.

Schneider, P., Biehl, M., & Hammer, B. (2009). Adaptive relevance matrices in learning vector quantization. *Neural Computation, 21*, 3532–3561.

Sillevis Smit, J., van Everdingen, J., Starink, T., & van der Horst, H. (Eds.) (2009). Dermatovenereologie voor de eerste lijn. Bohn Stafleu Van Loghum.

Stanganelli, I., Brucale, A., Calori, L., Gori, R., Lovato, A., Magi, S., et al. (2005). Computer-aided diagnosis of melanocytic lesions. *Anticancer Research, 25*, 4577–4582.

Suárez-Cuenca, J. J., Guo, W., & Li, Q. (2011). Automated detection of pulmonary nodules in CT: False positive reduction by combining multiple classifiers. In *Proc. SPIE* (Vol. 7963, pp. 796338–796338-6) http://dx.doi.org/10.1117/12.878793.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*, 411–423. http://dx.doi.org/10.1111/1467-9868.00293.

Walther, D., & Koch, C. (2006). 2006 special issue: Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407. http://dx.doi.org/10.1016/j.neunet.2006.10.001.

Zagrouba, E., & Barhoumi, W. (2011). A preliminary approach for the automated recognition of malignant melanoma. *Image Analysis & Stereology, 23*, 121–135<http://www.ias-iss.org/ojs/IAS/article/view/759>.