Learning Methods for Melanoma Recognition

Elisabetta La Torre,[1] Barbara Caputo,[2] Tatiana Tommasi[2]†

[1] Dipartimento di Fisica, Università di Roma "La Sapienza," Roma, Italy

[2] IDIAP Research Institute, Martigny, Switzerland

Received 23 July 2008; accepted 28 September 2010

ABSTRACT: Melanoma is the most deadly skin cancer. Early diagnosis is a challenge for clinicians. Current algorithms for skin lesions' classification focus mostly on segmentation and feature extraction. This article instead puts the emphasis on the learning process, testing the recognition performance of three different classifiers: support vector machine (SVM), artificial neural network and *k*-nearest neighbor. Extensive experiments were run on a database of more than 5000 dermoscopy images. The obtained results show that the SVM approach outperforms the other methods reaching an average recognition rate of 82.5% comparable with those obtained by skilled clinicians. If confirmed, our data suggest that this method may improve classification results of a computer-assisted diagnosis of melanoma. © 2010 Wiley Periodicals, Inc. Int J Imaging Syst Technol, 20, 316–322, 2010; Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/ima.20261

Key words: melanoma recognition; computer assisted diagnosis; dermoscopy; support vector machines; kernel methods

I. INTRODUCTION

Cutaneous melanoma is a spreading disease in the western world. As advanced melanoma is still practically incurable, early recognition and surgical excision of thin lesions remain the mainstay of treatment (Burroni et al., 2004). Despite the increasing awareness of melanoma (Rigel et al., 2000), clinical diagnostic accuracy is still disappointing (Burroni et al., 2004). Physicians visually inspect dermoscopic images for abnormal morphological and chromatic features that indicate malignancy. They commonly use the asymmetry, border, color, dimension, and dermoscopic (ABCD) structures rule for dermoscopy as guideline. Because of the subjective nature of examination, the accuracy of diagnosis is highly dependent on physician's expertise. Computer-aided diagnosis (CAD) system could provide an objective second opinion to clinicians, based on consistently extracting and analyzing image features (Burroni et al., 2004). The topic is largely investigated (Ganster et al., 2001; Rubegni et al., 2002; Grana et al., 2003; Sboner et al., 2003; Schmid-Saugeon et al., 2003; Maglogiannis et al., 2005; Celebi et al., 2007). The mainstream approach to the problem focuses on the development of segmentation algorithms (Ganster et al., 2001; Grana et al., 2003; Schmid-Saugeon et al., 2003; Celebi et al., 2007) and ad hoc feature descriptors (Ganster et al., 2001; Rubegni et al., 2002; Grana et al., 2003; Sboner et al., 2003; Maglogiannis et al., 2005; Celebi et al., 2007). The segmentation step is important because it separates the lesion from the surrounding skin and hairs (Ganster et al., 2001; Schmid-Saugeon et al., 2003; Celebi et al., 2007), and it extracts the lesion's contour, a relevant component in the diagnostic process (Grana et al., 2003). Equally important is the choice of the features that correspond to the selection of the relevant information for the final diagnosis (Rubegni et al., 2002; Grana et al., 2003; Sboner et al., 2003; Maglogiannis et al., 2005; Celebi et al., 2007). The classification algorithm, the final ingredient in any CAD system, is typically taken from the pattern recognition literature. Examples of classification algorithms used in the last years are *k*-nearest neighbors, *k*-NNs (Ganster et al., 2001; Sboner et al., 2003), artificial neural networks, ANNs (Rubegni et al., 2002; Maglogiannis et al., 2005), and very recently support vector machines, SVMs (Celebi et al., 2007).

The contribution of this article is a comparative evaluation of several learning methods on a large collection of skin lesion images. Specifically, we selected three different classifiers: SVMs (Vapnik, 1998), ANNs (Bishop, 1995), and *k*-NN (Bishop, 1995). We conducted an experimental evaluation of these techniques on the Ganster's database,* a collection of more than 5300 skin lesion dermoscopy images. Using this database permits to compare our results with those of expert clinicians and with the Ganster's method, based mainly on sophisticated segmentation and feature extraction algorithms. We tested the classification methods on two different types of features, color histograms, CHs (Swain et al., 1991), and multidimensional receptive fields histograms, MFHs (Schiele et al., 2000). These features reproduce two of the criteria followed by dermatologists for diagnosis, respectively, ''C'' for color variegation and ''D'' for differential local structures. Our results show that SVM obtains remarkably better performances than all other considered methods. It is remarkable to note that, on two classes out of three, SVM achieves recognition results comparable with those obtained by skilled clinicians. The rest of the article is organized as follows: Section II describes the database used.

†Work done while at ''Università di Roma La Sapienza''.

Correspondence to: Elisabetta La Torre; e-mail: elisabetta.latorre@uniroma1.it

*We gratefully thank H. Ganster and A. Pinz for making the database and their segmentation masks available to us.

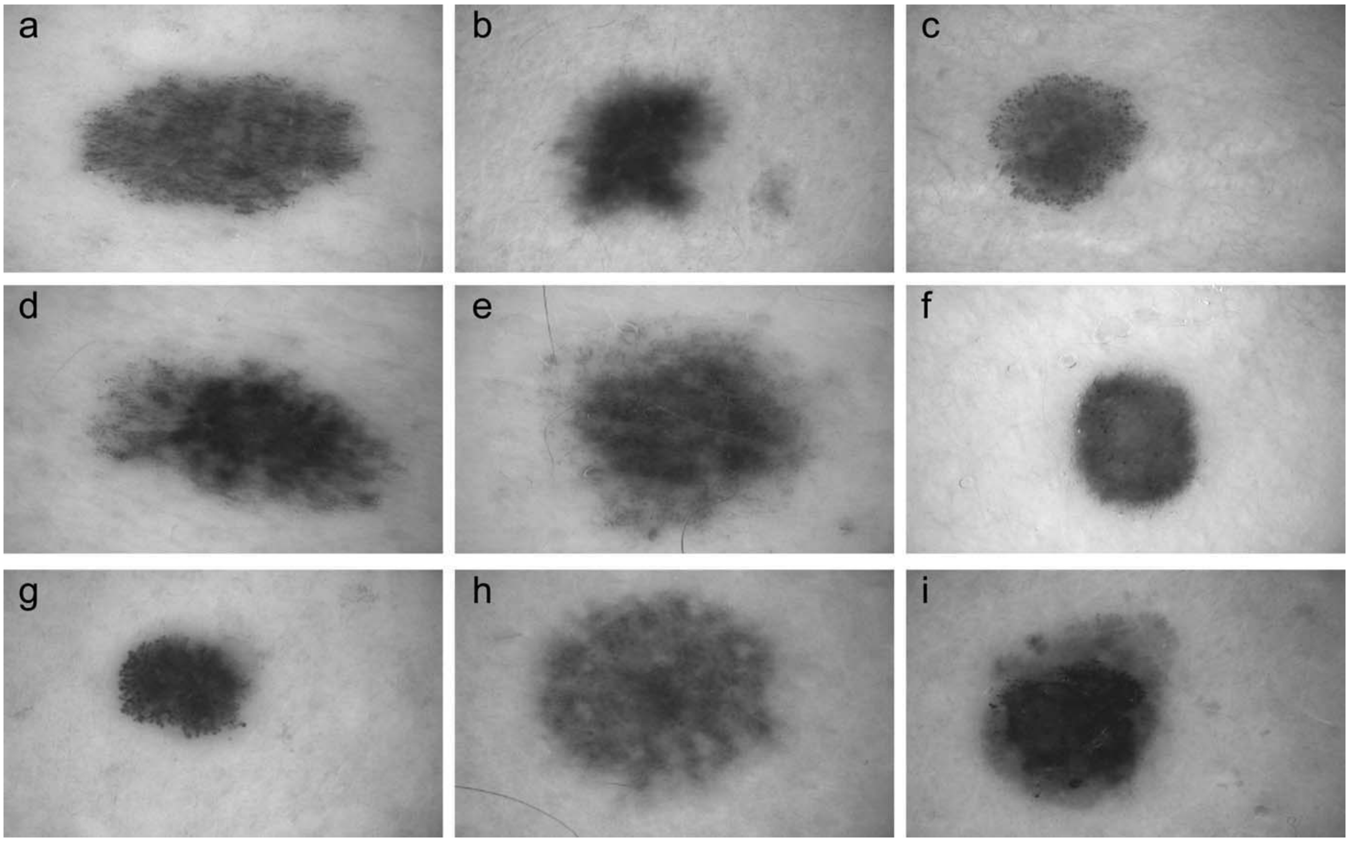

Figure 1. Examples of skin lesion's images used: images of (a, b, and c) benign lesions, (d, e, and f) dysplastic lesions, and (g, h, and i) malignant lesion. Note the high variability within one class and the low variability between different classes. This makes the classification problem very challenging.

Section III briefly reviews the segmentation procedure and features extraction algorithms chosen in the article. Section IV describes our classification methods and Section V reports our experimental findings. The article concludes with a summary discussion and possible directions for future research.

II. DATASET DESCRIPTION

We performed our experiments on the database created by the Department of Dermatology of the Vienna General Hospital (Ganster et al., 2001). The authors refer that all images were captured during routine clinical examinations to reflect the a priori probabilities of the routine diagnosis in a specialized dermatology clinic (Ganster et al., 2001). The whole database consists of 5380 skin lesion images, divided into three classes: 4277 of these lesions are classified as clearly benign lesions (class 1), 1002 are classified as dysplastic lesions (class 2), and 101 lesions are classified as malignant melanomas (class 3).[†] The lesions of the classes 2 and 3 were all surgically excised, and the ground truth was generated by means of histological diagnosis (Ganster et al., 2001). To have statistically significant results, we ran experiments with five different partitions, selected at random. More details about the partitions' selection are explained in Section V. This procedure has been adopted for all the experiments reported here. Figure 1 shows some exemplar images for each class.

III. SEGMENTATION AND FEATURE EXTRACTION

A. Preprocessing and Segmentation. Following the approach proposed in the study of Ganster et al. (2001), we did not implement any preprocessing step such as color normalization or hair removal. As for the segmentation procedure, we used two different methods. The first consists of simply cutting all the images with the help of a common image editor software, selecting for each image the smallest rectangle containing the lesion and keeping out as much skin as possible. We call the resulting images ''hand-segmented''. The second method is the one developed by Ganster et al. (2001). It consists of a binary mask determined by several segmentation algorithms combined together with a fusion strategy. We call the resulting images ''mask-segmented.'' Exemplar images obtained by these two segmentation techniques are shown in Figure 2. Running experiments on these two types of images allow us to explore how the classification performance is affected by the quality of the segmentation process.

B. Feature Extraction. In the ABCD rule, the color variegation and the dermoscopic structures in the skin lesion are two of the discriminant characteristics for clinical melanoma recognition. Thus we decided to use CH and MFH as features able to retain chromatic and textural information, respectively. A CH denotes the joint probabilities of the intensities of the three color channels (Swain et al., 1991). The red, green, and blue (RGB) system is based on the fact that a large percentage of the visible spectrum can be represented by mixing RGB colored light in various proportions and intensities.

[†] These numbers are not perfectly coincident with those reported in the study of Ganster et al. (2001), where the database is said to be of 5363 images, but this difference should not affect the comparison between the two algorithms.

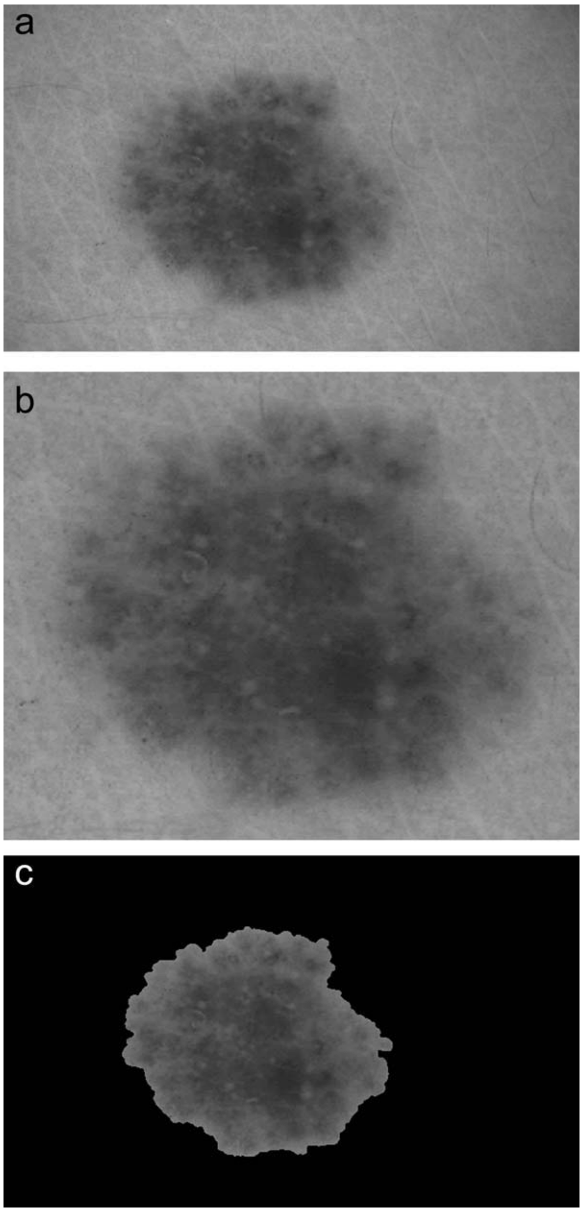

Figure 2. Examples of segmentation. (a) An example of an entire image, (b) the same image hand-segmented, and (c) the same image mask-segmented.

If the red and green components are normalized with the sum of the RGB components, the rg system is obtained. Another common approach is using the hue–saturation–value (HSV). This takes into account the perceptual differences of colors by describing them by their perceived color (hue), their dilution by white light (saturation), and their luminance values. A CH of an image is produced first by discretization of the colors in the image into a number of bins and counting the number of image pixels in each bin. For each series of experiments we used hue, rg, RG, RB, and GB CHs. Also, the resolution of the bin axes was varied for each representation, consisting of 8, 16, 32, and 64 (for bidimensional histograms we chose the resolution of each axis with the same bin value). For space reasons, we report only the best result obtained.

The main idea of MFH is to calculate multidimensional histograms of the response of a vector of receptive fields (Schiele et al., 2000). A MFH is determined once we chose the local property measurements (i.e., the receptive field functions), which determine the dimensions of the histogram and the resolution of each axis. We converted originally RGB images to gray scale and then we used two different kinds of MFH representations. The first consisted of Gaussian derivatives along x and y directions (Schiele et al., 2000):

$$G_x^\sigma(x,y) = -\frac{x}{a^2}G^\sigma(x,y), \quad G_y^\sigma(x,y) = -\frac{y}{a^2}G^\sigma(x,y) \tag{1}$$

where $G^\sigma(x,y)$ is the Gaussian distribution (Schiele et al., 2000) with $\sigma = 1.0(DxDy)$. The second consisted of Laplacian Gaussian operator (Schiele et al., 2000):

$$Lap(x,y) = \left(\frac{x^2}{\sigma^4}+\frac{1}{\sigma^2}\right)G^\sigma(x,y) + \left(\frac{y^2}{\sigma^4}+\frac{1}{\sigma^2}\right)G^\sigma(x,y) \tag{2}$$

where $G^\sigma(x,y)$ is the Gaussian distribution (Schiele et al., 2000) with $\sigma_1 = 1.0$, 1.5, and 3.0 and $\sigma_2 = 2.0$, 3.0, and 6.0 respectively ($Lp2\sigma$). The bin axes' resolution varied for each representation. It was of 8, 16, 32, and 64 for $DxDy$ and 16 and 32 for $Lp2\sigma$.

IV. CLASSIFICATION

In this section, we describe the classification algorithms that we compared in this article: SVMs, ANNs, and k-NNs.

A. SVMs. SVMs are state-of-the-art large margin classifiers. Here, we provide a brief review of the theory behind this type of algorithm for the two class case. For the extension to multiclasses and a more detailed treatment, we refer to the study of Vapnik (1998). Consider the problem of separating a set of training data $(x_1,y_1), \ldots, (x_m,y_m)$, where $x_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$, its class label. If we assume that the two classes can be separated by a hyperplane $\boldsymbol{w}\cdot\boldsymbol{x} + b = 0$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane is that which has maximum distance to the closest points in the training set. The optimal values for $\boldsymbol{w}$ and b can be found by solving a constrained minimization problem, resulting in the classification function

$$f(x) = sign\left(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{w}\cdot\boldsymbol{x} + b\right) \tag{3}$$

where α_i and b are found during training (Vapnik, 1998). Most of the α_i's take the value of zero; those x_i with nonzero α_i are the ''support vectors.'' In case where the two classes are not linearly separable an upper bound on the Lagrange multipliers is introduced $\alpha_i \leq C$, $i = 1, \ldots, m$, where C determines the trade-off between margin maximization and training error minimization. It is also possible to give different costs to false-positive and false-negative errors, introducing the parameters C^+ and C^-, respectively, instead of C (Vapnik, 1998). To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a high-dimensional feature space H by $x \rightarrow \phi(x)$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(x,y) = \phi(x)\cdot\phi(y)$, a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{w}\cdot\boldsymbol{x}$ by the kernel function in Eq. (3). Popular kernel functions are the Gaussian radial basis function (RBF) kernel

$$k(x,y) = \exp(-\gamma^*|x-y|^2) \tag{4}$$

and the polynomial kernel

$$k(x, y) = (\gamma^* x \cdot y)^d \tag{5}$$

As SVMs have started to be used for visual recognition, several researchers have proposed new kernel functions. Chapelle et al. (1999) proposed two new types of exponential kernels: the generalized Gaussian RBF kernel

$$k(x, y) = \exp(-\gamma^* |x^a - y^a|^b) \tag{6}$$

and the χ^2 kernel.

$$k(x, y) = \exp(-\gamma^* \chi^2(x, y)) \tag{7}$$

These are the four kernels used in this article.

B. ANNs. ANNs are learning approaches inspired by the way biological nervous systems process information. They are composed of a large number of highly interconnected elements designed to mimic neurons (Bishop, 1995). An artificial neuron receives a number of inputs either from original data or from the output of the other neurons in the network. Each input comes via a connection that has a strength, which corresponds to synaptic efficacy. Each neuron has also a single threshold value. The weighted sum of the inputs is calculated and the threshold subtracted to compose the activation signal of the neuron, which is then passed through a transfer function to produce the output response. Typical transfer functions are linear, threshold, and sigmoid (Bishop, 1995). Neural networks are usually organized in layers: the first one is called the input layer; the last one, the output layer; the intermediate ones (if any) are called the hidden layers (Abdi et al., 1999). In a feed-forward network, the information travels one way only, from input to output, whereas in feedback networks, signals travel in both directions by introducing loops. The learning process is achieved through the modification of the connection weights and thresholds between units. The most widely known supervised learning rule uses the difference between the actual output of a cell and the desired output as an error signal for units in the output layer. The adaptation of this rule for a multilayer network is known as error backpropagation (Abdi et al., 1999). In this article, we used a feed-forward neural network with only one hidden layer, sigmoid transfer function, and error backpropagation as learning rule.

C. *k*-NN. The *k*-NN algorithm is a classification method based on closest training examples in the feature space. Let $D^n = \{x_b, \ldots, x_n\}$ denote a set of *n* labelled prototypes and letting $x' \in D^n$ be the prototype nearest to a test point $\boldsymbol{x}$. Then the nearest-neighbor rule for classifying x is to assign it the label associated with x' (Bishop, 1995). An obvious extension of the nearest-neighbor rule is the *k*-NN rule. This rule classifies $\boldsymbol{x}$ by assigning it the label most frequently represented among the *K*-nearest samples (Bishop, 1995). The *k*-NN query starts at the test point $\boldsymbol{x}$ and grows as a spherical region until it encloses *k* training samples, and it labels the test point by a majority vote of these samples (Bishop, 1995). If *k* is fixed and the number *n* of samples is allowed to approach infinity, then all of the *k*-NNs will converge to $\boldsymbol{x}$ (Bishop, 1995).

V. EXPERIMENTS

In this section, we present experiments that show the effectiveness of SVMs for melanoma recognition. We bench-marked SVMs, ANN, and *k*-NN. For SVM, we used the four kernel types described in Section IV. The kernel parameters were chosen via cross-validation. For ANN, we used a multilayer perceptron (MLP) with normalized features. We tested MLP with one hidden layer, varying the number of hidden neurons, i.e., h = 514, 257, 128, and 64. For the *k*-NN experiments, we used a normalized Euclidean distance and K = 1, 3, 5, . . ., 29. For space reasons, we report only the best results obtained for ANN and *k*-NN in the following sections.

Table I. Recognition results for the first series of experiments

	CH, hand	CH, mask	MFH, hand	MFH, mask
SVM-Poly (%)	74.9 ± 2.8	67.1 ± 7.8	66.9 ± 13.1	69.7 ± 3.8
SVM-Gauss (%)	59.0 ± 10.3	62.6 ± 6.2	51.1 ± 8.6	69.8 ± 3.7
SVM-genGauss(%)	75.9 ± 14.0	80.2 ± 2.8	80.7 ± 1.7	82.5 ± 0.1
SVM-χ (%)	76.0 ± 13.7	59.9 ± 12.9	70.3 ± 1.5	81.0 ± 0.2
ANN (%)	55.8 ± 4.7	64.0 ± 10.1	51.2 ± 12.8	48.0 ± 9.8
K-NN (%)	46.0 ± 5.4	48.9 ± 5.5	47.1 ± 4.9	51.4 ± 4.7
Ganster et al. (%)	58			

The experiments were performed on hand-segmented and mask-segmented images using SVM with four kernel types, ANN and *K*-NN, with CH and MFH representation as features. Results are averaged on five partitions, and standard deviations are also calculated. The results obtained by Ganster et al. (2001) are also reported; note that these results were obtained on a single run.

We performed three series of experiments: in the first series, all the experiments were performed respecting the procedure reported by Ganster et al. (2001). The training set consisted of 270 images; those images were selected at random by choosing five sets of 90 images for each class. For classes 1 and 2, it was also possible to impose the condition to have five different and disjoint training sets. This constraint was not applied on class 3 because of the few number of images, so the obtained five sets for this class were not disjoint. The number of images in class 3, that is the images which present the disease, depends on the prevalence of the disease itself. The test set consisted of the whole database (Ganster et al., 2001). Note that training and test set are not disjoint; once again we underline that this follows the procedure proposed by Ganster et al. (2001), allowing for comparison of results. The outcome of these experiments is reported in Section VA. For a fair evaluation of the learning algorithms, it is necessary to disjoin the training set from the test set. Therefore, we performed a second series of experiments using the following partition: the training set consisted of 270 images (90 for each class); those images were selected following the same method chosen for the first series of experiments. The test set consisted of the remaining database. The results of these experiments are reported in Section VB. Finally, a third series of experiments was performed, using only SVM for binary classification. Indeed, if the aim of a CAD for skin lesions classification is to prescribe or not the surgical excision of the lesion, it is reasonable to group dysplastic and malignant lesions into a common class. In this series of experiments, the database was thus composed of two classes: the first coincident with class 1 of the previous experiments, and the second defined by the union of classes 2 and 3 of the previous experiments. The training set consisted of 180 images (90 for each of the two obtained classes); the test set consisted of the remaining database. The results of these experiments are reported in Section VC.

A. First Series of Experiments. We performed experiments using CH and MFH representations as features. The obtained recognition rates for hand-segmented and mask-segmented images using SVM, ANN, and *k*-NN with both features types are reported in Table I. For sake of clarity, we report the results obtained by Ganster et al. (2001) too. Note that these results were obtained on a single run, and as the features used are very different from ours, the comparison between the approaches is mostly indicative.

A first comment is that SVM obtains the best result with respect to Ganster's method, ANN, and *k*-NN, for both feature types and for both segmentation strategies. The best result is of 82.5%, obtained using the generalized Gaussian kernel, MFH features, and mask-segmented images. Comparable results are obtained with color features, selected kernels, and on hand-segmented images. The best performance achieved by ANN is of 64.0%, obtained using mask-segmented images and CH features, with 257 neurons in the hidden layer. Finally, the best result achieved by *k*-NN is of 51.4%, obtained using mask-segmented images and MFH features, with $K = 3$. The recognition rate obtained with the Ganster's method is of 58%. These results clearly suggest the effectiveness of SVMs for melanoma recognition. A second comment is that SVM's performance varies considerably depending on the kernel type used. For instance, using color features and hand-segmented images, the recognition rate goes from a minimum of 59.0% for the Gaussian kernel to a maximum of 76.0% for χ^2-kernel. A similar behavior is observed by using mask-segmented images and on textural features. It is also interesting to note that, with both segmentation techniques and feature types, the kernels which obtain the worst performances tend to have the highest standard deviations, whereas the kernel with the best performance has the smallest one. This illustrates the importance of doing kernel selection during training; the low standard deviation of the SVMs' best results also shows the stability of our findings. We observe that the results obtained by the polynomial kernel are comparable with those given by the other exponential kernels. As the polynomial kernel is computationally very expensive, we decided not to use it in the rest of the experiments. By comparing the hand-segmented best result with the mask-segmented one, we can see an improvement in recognition rate and stability passing from the first to the second, for both feature types. This is an experimental proof of the importance of using a sophisticated segmentation method.

Table II reports the confusion matrices for the best results obtained by each possible combination of segmentation mask, feature type and SVMs, plus the confusion matrix obtained by Ganster and that relative to clinicians' performance on the database (Ganster et al., 2001).[‡] For both segmentation techniques and feature types, we see that SVM outperforms Ganster's method for classes 1 and 3, and it is comparable with the dermatologists' performances. It is very interesting to note that, in contrast, SVM performs poorly on class 2, which corresponds to dysplastic lesions. This might be explained considering that here we are using only one feature type for each set of experiments, whereas Ganster used a selection of different features, and dermatologists used the ABCD rule. It is thus possible that just color/textural information is not discriminant enough to recognize correctly dysplastic lesions, whereas both feature types seem to be effective for separating benign and malignant lesions.

Table II. Confusion matrices for the first series of experiments

Ganster et al.				Clinicians			
	Assigned				Assigned		
True	Class 1	Class 2	Class 3	True	Class 1	Class 2	Class 3
Class 1	2500	1347	410	Class 1	4161	94	9
Class 2	324	531	155	Class 2	42	960	8
Class 3	14	12	70	Class 3	6	19	78
	SVM, CH hand				SVM, CH mask		
Class 1	3850.6	259.4	167.0	Class 1	4112.6	112.6	50.8
Class 2	798.2	150.4	53.4	Class 2	874.8	110.0	17.2
Class 3	9.8	1.2	90.0	Class 3	10.4	0.2	90.4
	SVM, MFH hand				SVM, MFH mask		
Class 1	4184.8	45.5	45.8	Class 1	4251.8	4.2	20.0
Class 2	861.6	116.8	23.6	Class 2	901.0	95.8	5.2
Class 3	9.8	0.6	90.6	Class 3	10.4	0.4	90.2

Results for each class are averaged on five partitions, and standard deviations are also calculated. The confusion matrix obtained by Ganster et al. (2001) and that relative to clinicians' performance on the database (Ganster et al., 2001) are also reported.

Table III. Recognition results for the second series of experiments

	CH, hand	CH, mask	MFH, hand	MFH, mask
SVM-Gauss (%)	58.8 ± 11.4	62.8 ± 7.1	51.0 ± 9.26	70.8 ± 4.0
SVM-genGauss (%)	75.0 ± 13.8	79.1 ± 3.0	79.5 ± 1.8	82.9 ± 0.9
SVM-χ (%)	74.8 ± 14.3	60.2 ± 12.3	68.8 ± 1.6	78.8 ± 1.5
ANN (%)	55.8 ± 5.2	64.6 ± 11.2	50.6 ± 13.6	44.9 ± 6.5
K-NN (%)	44.4 ± 5.7	47.5 ± 5.9	45.9 ± 5.0	50.3 ± 4.9

The experiments were performed on hand-segmented and mask-segmented images using SVM with three kernel types, ANN and *K*-NN, with CH, and MFH representation as features. Results are averaged on five partitions, and standard deviations are also reported.

B. Second Series of Experiments. Experiments reported in Section VA were performed on a not-disjoint experimental set. This was done to compare fairly with the results reported by Ganster et al. (2001), but this strategy does not allow to evaluate properly the generalization capability of the chosen learning methods. Thus, we performed a second series of experiments using a disjoint training and test set partitioned as follows: the training set consisted of 270 images (90 for each class); the test set consisted of the remaining database. As in the previous series of experiments, we used CH and MFH as features. For classification, we used SVM with kernel functions χ^2, the Gaussian, and the generalized Gaussian ones, ANN and *k*-NN. The classification results for hand-segmented and mask-segmented images, with both features types and for all the learning methods used, are shown in Table III.

We see that the best recognition rate is of 82.9 ± 0.9%, obtained using the generalized Gaussian kernel, MFH features, and mask-segmented images. This result must be compared with the best recognition rate obtained on training and test set not-disjoint (Section VA and Table I), which was of 82.5 ± 0.1%. Both these results were obtained using the generalized Gaussian kernel, MFH features, and mask-segmented images. These two results are statistically equivalent and confirm the suitability of SVM for this application. As we noted earlier, SVM's performance varies considerably depending on the kernel type used. Once again the best performance is achieved with the generalized Gaussian kernel, for all the feature representations and for both hand-segmented and mask-segmented images. On the basis of the results obtained on these two first series of experiments, we conclude that SVMs are the best classification method among those proposed here for melanoma classification. For the last series of experiments, we will therefore use only SVM with different kernel functions.

C. Two Class Experiments and Receiver Operating Characteristic Analysis. As reported by Ganster et al. (2001), during routine clinical practice in the Vienna General Hospital, the

[‡] For more details on the number of images used in the these last two confusion matrices we refer the reader to Ganster et al. (2001).

lesion is not surgically excised if three experienced dermatologists agree on a benign diagnosis (Ganster et al., 2001). The lesions named dysplastic are still considered as benign, but are so-called precursors of malignant melanoma. As this category represents skin lesions with an increased risk to turn into a melanoma, the category receives its own class label. The lesions classified as dysplastic are all surgically excised, as it is done for malignant ones (Ganster et al., 2001).

Given that the purpose of a CAD system is to prescribe or not the surgery, we decided to evaluate the recognition performance of SVM for the classification of skin lesions in two classes: the first class consisting of images of clearly benign lesions, and corresponding to class 1; the second class consisting of the union of dysplastic and malignant lesion images, so we have now a new ''class 2'' which is constituted by the union of the old classes 2 and 3. Thus we expect the CAD system to suggest the surgical excision for the lesions in this last class on the basis of a binary decision. We thus performed a new series of experiments with the following experimental setup: we used the ''mask-segmented'' images and CH and MFH as features. The training set consisted of 180 images, 90 for each class, to have the same number of images in the training set for the two new classes; the test set consisted of the remaining database. We used SVM with the χ^2, Gaussian, and generalized Gaussian kernels. The kernel parameters were chosen via cross-validation; the obtained results were analyzed using the receiver operating characteristic (ROC) analysis (Van Erkel et al., 1998). Specifically, we posed the cost parameter (C parameter, Section IV) equal to 1, and we varied the C^+/C^- ratio (see Section IV) from one to nine to obtain the different points of the ROC curve. This means that the loss of true positives is weighted more and more as the C^+/C^- ratio increases. We then used the area under the ROC curve (AUC) as a summary measure of the overall diagnostic performance (Hanley et al., 1982). Table IV reports the average values of sensitivity and specificity, with their standard deviations, for each kernel and for each feature type, for $C = 1$ and $C^+/C^- = 1$. Table V reports the average values of the AUC and their standard deviations for each kernel.

These results clearly show that SVM gives very high values in sensitivity with χ^2 and generalized Gaussian kernels and CH and with generalized Gaussian kernel and MFH. In particular, with generalized Gaussian kernel, we have a surprising sensitivity of 100% with a null standard deviation for both feature types. It means that all the positive lesions are correctly classified within every partition and with both the features representation. Generalized Gaussian kernel gives the best specificity also with the lowest standard deviations; for this kernel we have a 99.23% for CH and 99.47% for MFH.

VI. CONCLUSIONS

In this article, we evaluated the importance of the classification method for melanoma recognition. To this purpose, we considered three different classifiers: SVMs, ANNs, and k-NN. The classifiers were tested on a database of more than 5000 images using two feature types and two segmentation methods. Our results show that SVM achieves very high performance on this task compared with the other learning methods and compared with a feature-based method previously proposed in the literature (Ganster et al., 2001). Moreover, on two classes out of three, SVM achieves recognition results comparable with those obtained by skilled clinicians. A series of two class experiments showed that SVM gives very good results in sensitivity and specificity, suggesting that SVM could be an aid for the physicians in the choice of prescribing or not the surgical excision of the lesion. Data should be confirmed in an unselected population of cases representative of a true clinical context, including different kinds of lesions such as junctional, Spitz nevi, or other common nevi. In the future, we plan to conduct similar experiments using shape descriptors, and finally to experiment with cue integration schemes, to test the effectiveness of different types of information and eventually to reproduce the ABCD method.

Table IV. Sensitivity and specificity results for the third series of experiments

	Sensitivity		Specificity	
Kernel	CH (%)	MFH (%)	CH (%)	MFH (%)
χ	99.78 ± 0.11	77.87 ± 8.59	76.46 ± 7.37	97.97 ± 0.75
Gauss	64.76 ± 19.69	85.88 ± 1.35	85.70 ± 3.80	57.42 ± 9.51
GenGauss	100.00 ± 0.00	100.00 ± 0.00	99.23 ± 0.25	99.47 ± 0.35

Results are averaged on five partitions, and standard deviations are also reported.

Table V. AUC average values for the third series of experiments

	AUC	
Kernel	CH features	MFH features
χ	0.882 ± 0.037	0.987 ± 0.005
Gauss	0.832 ± 0.039	0.741 ± 0.052
Gengauss	0.995 ± 0.003	0.997 ± 0.002

Results are averaged on five partitions, and standard deviations are also reported.

REFERENCES

H. Abdi, V. Valentin, and B. Edelman, Neural networks, Sage, Thousand Oaks, CA,1999. Available at: citeseer.ist.psu.edu/abdi03neural.html

C.M. Bishop, Neural networks for pattern recognition, Claredon Press, Oxford,1995.

M. Burroni, R. Corona, G. Dell'Eva, F. Sera, R. Bono, P. Puddu, R. Perotti, F. Nobile, L. Andreassi, and P. Rubegni, Melanoma computer-aided diagnosis: Reliability and feasibility study. Clin Cancer Res 10 (2004), 1881–1886.

M.E. Celebi, H.A. Kingravi, B. Uddin, H. Iyatomi, Y.A. Aslandogan, W.V. Stoecker, and R.H. Moss, A methodological approach to the classification of dermoscopy images, Comput Med Imaging Graph 31 (2007), 362–373.

O. Chapelle, P. Haffner, and V. Vapnik, SVMs for histogram based image classification, IEEE Trans Neural Netw 10 (1999), 1055–1064.

H. Ganster, A. Pinz, R. Roddothrer, E. Wildling, M. Binder, and H. Kittler, Automated melanoma recognition, IEEE Trans MI 20 (2001), 233–239.

C. Grana, G. Pellacani, R. Cucchiara, and S. Seidenari, A new algorithm for border description of polarized light surface microscopic images of pigmented skin lesions, IEEE Trans MI 22 (2003), 959–964.

J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982), 29–36.

I. Maglogiannis, S. Pavlopoulos, and D. Koutsouri, An integrated computer supported acquisition, handling, and characterization system for pigmented skin lesions in dermatological images, IEEE Trans Inf Technol Biomed 9 (2005), 86–98.

D.S. Rigel and J.A. Carucci, Malignant melanoma: Prevention, early detection, and treatment in the 21st century, CA Cancer J Clin 50 (2000), 215–236.

P. Rubegni, M. Burroni, G. Cevenini, R. Perotti, G. Dell'Eva, P. Barbini, M. Fimiani, and L. Andreassi, Digital dermoscopy analysis and artificial neural network for the differentiation of clinically atypical pigmented skin lesions: A retrospective study, J Invest Dermatol 119 (2002), 471–474.

A. Sboner, C. Eccher, E. Blanzieri, P. Bauer, M. Cristofolini, G. Zumiani, and S. Forti, A multiple classifier system for early melanoma diagnosis, Artif Intell Med 27 (2003), 29–44.

B. Schiele and J.L. Crowley, Recognition without correspondence using multidimensional receptive field histograms, Int J Comput Vis 36 (2000), 31–2.

P. Schmid-Saugeon, J. Guillod, and J.P. Thiran, Towards a computer-aided diagnosis system for pigmented skin lesions, Comput Med Imaging Graph 27 (2003), 65–78.

M. Swain and D. Ballard, Color indexing, Int J Comput Vis 7, 1991, 11–32.

A.R. Van Erkel and P.M.T. Pattynama, Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology, Eur J Radiol 27 (1998), 88–94.

V. Vapnik, Statistical learning theory, Wiley, New York, 1998.