IDIAP RESEARCH REPORT

Melanoma Recognition using Kernel Classifiers

Elisabetta La Torre [a] Barbara Caputo [b c]
Tatiana Tommasi [a]

IDIAP–RR 06-53

October 2006

submitted for publication

[a] University of Rome "La Sapienza", Physics Department
[b] IDIAP - bcaputo@idiap.ch
[c] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-53

Melanoma Recognition using Kernel Classifiers

Elisabetta La Torre Barbara Caputo Tatiana Tommasi

October 2006

submitted for publication

Abstract. Melanoma is the most deadly skin cancer. Early diagnosis is a current challenge for clinicians. Current algorithms for skin lesions classification focus mostly on segmentation and feature extraction. This paper instead puts the emphasis on the learning process, proposing two kernel-based classifiers: support vector machines, and spin glass-Markov random fields. We benchmarked these algorithms against a state-of-the-art method on melanoma recognition. We show with extensive experiments that the support vector machine approach outperforms the other methods, proving to be an effective classification algorithm for computer assisted diagnosis of melanoma.

1 Introduction

Malignant melanoma is a spreading disease in the western world. Its incidence has been increasing over the past decades; currently 132,000 melanoma skin cancer occurs globally each year. One in every three cancers diagnosed is a skin cancer and, according to Skin Cancer Foundation Statistics, one in every five Americans will develop this kind of tumor during their lifetime [1]. Management of melanoma is a complex issue requiring a multidisciplinary approach. The most effective method of protection against the development of skin cancer is minimization of ultraviolet exposure from sunlight. Since advanced melanoma is still practically incurable, early recognition and surgical excision of thin lesions remains the mainstay of treatment, because tumor thickness is universally recognized as the primary determinant of prognosis [2]. Despite the increasing awareness of melanoma, because of the worldwide increase of incidence reported in the last few decades [3], clinical diagnostic accuracy is still disappointing [2]. Subsequent attempts to develop non-invasive tools to improve early diagnosis resulted in two approaches: EpiLuminescence Microscopy (ELM or dermoscopy) and digital image analysis. ELM allows the examination of a suspicious skin lesion by using an hand-held device emitting incident light from a light source penetrating the epidermal skin layer. It has been reported to improve the accuracy in diagnosing cutaneous lesions, including melanoma [2]. Physicians visually inspect dermoscopic images for abnormal morphologic and chromatic features that indicate malignancy. They commonly use the ABCD (Asymmetry, Border, Color, Dimension and Dermoscopic structures) method as guideline. Due to the subjective nature of examination, the accuracy of diagnosis is highly dependent upon physician's expertise.

There is a growing awareness that one of the weakest links in the biomedical interpretation process is the perception of details, and the recognition of their meaning by dermatologists. Computer Aided Diagnosis (CAD) system could provide to clinicians an objective second opinion, based on consistently extracting and analyzing image features [2]. Such a system should reproduce the perceptual and cognitive strategy followed by doctors, and should allow the dermatologist to trace each step of the process which led to a given diagnosis, so to leave space for exploring multiple interpretations. Recently numerous research on this topic have been proposed (for a more comprehensive discussion of the most significant literature we refer the reader to section 2). A key factor for the development and evaluation of these systems is the availability of a statistically significant database. One of the largest databases of melanoma images available to the research community was contributed by H. Ganster et al. [4]. In that paper they presented a database of 5363 images, accompanied by: (a) a segmentation algorithm for isolating the potential melanoma from the surrounding skin, determined by several basic segmentation algorithms combined together with a fusion strategy [4]; (b) a set of features containing shape and radiometric features as well as local and global parameters, calculated to describe the malignancy of a lesion, from which significant features are selected by application of statistical methods for feature selection [4]; (c) a nearest neighbor classification algorithm [4]. In that work the authors concentrated particularly on the segmentation technique and the feature selection process, obtaining results that, to the best of our knowledge, represent the state-of-the-art on this topic. Here we focus instead on the classification algorithm, proposing to use kernel methods for classification of skin lesion images. Specifically, we selected a discriminative method and a probabilistic one. As discriminative method we chose Support Vector Machines (SVM, [5]), a state-of-the-art large margin classifier, where the optimal separating surface is defined by a linear combination of scalar products between the view to be classified and some support vectors [6][5]. By introducing a Mercer kernel, a non-linear SVM can be constructed replacing the scalar products in the linear SVM via the kernel function. SVMs have demonstrated remarkable performance on object recognition and categorization [7]. As probabilistic method we chose Spin Glass-Markov Random Fields (SG-MRF, [8]), a fully connected MRF which integrates results of statistical mechanics with Gibbs probability distributions via non linear kernel mapping [8]. Experiments have shown the robustness and categorization capabilities of this algorithm for object recognition [8] and its applicability for biomedical applications [9]. We conducted an experimental evaluation of these two techniques on the Ganster's database[1], which al-

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

lows for a straightforward benchmarking of our algorithms against theirs. We tested our two methods on two different types of features, Color Histograms (CH, [10]) and Multidimensional receptive Fields Histograms (MFH, [11]). These features reproduce two of the criteria followed by dermatologists for diagnosis, respectively "C" for color variegation and "D" for differential local structures. We also evaluated the influence of the segmentation method by running two series of experiments: the first using the segmentation masks obtained by Ganster, the second using an hand-made rectangular mask which roughly contains the whole lesion while minimizing the amount of surrounding skin in the image. In order to have a fair comparison, we first replicated the experimental setup used in [4] for a benchmark evaluation. Then, we used different partitions of the database (with respect to training and test, and with repsect to the number of classes), so to reproduce more realistic scenarios. Our results show that SVM obtains remarkably better performances than SG-MRF and Ganster's method with both feature types, regardless of the segmentation method. It is remarkable to note that, on two classes out of three, SVM achieves recognition results comparable to those obtained by skilled clinicians.

The rest of the paper is organized as follows: section 2 reviews the state of the art in computer-assisted melanoma recognition. Then we briefly review the theory behind SG-MRF (section 3) and SVMs (section 4). Section 5 describes the experimental setup and reports on our findings. The paper concludes with a summary discussion and some possible directions for future research.

2 Related Work

Recently there has been an increasing interest in developing algorithms for melanoma classification. Grana et al. [12] provided mathematical descriptors for the border of pigmented skin lesion images, and assessed their efficacy for distinction among different lesion groups. They introduced new descriptors, such as lesion slope and lesion slope regularity, and defined them mathematically. Then, they employed a new algorithm based on the Catmull Rom spline method and the computation of the gray-level gradient of points extracted by interpolation of normal direction on spline points [12]. The efficacy of these descriptors was tested on a data set of 510 pigmented skin lesions, composed by 85 melanomas and 425 nevi, by employing statistical methods for discrimination between the two populations [12].

Maglogiannis et al. [13] described an integrated prototype system which includes an image acquisition arrangement, designed for capturing skin images under reproducible conditions. The system processes the captured images and performs unsupervised image segmentation and image registration utilizing an algorithm based on the log-polar transform of the images' Fourier spectrum [13]. Six algorithms for image segmentation were tested by the authors for their efficiency: thresholding, weighted functions, region growing, Principal Components Transform (PCT), CIELAB color space transform and spherical coordinates transform [13]. Border- and color-based features were computed. Neural networks and discriminant analysis were used for classification of malignant melanoma versus dysplastic nevus [13]. The database consisted of three groups of data: the first group consists of 14 cases of malignant melanoma, with measurements taken on the entire extent of the lesion. The second group also referred to the malignant melanomas, but measurements were restricted to the dark area of the melanoma. The third group contained 20 cases of dysplastic nevus.

Schmid-Saugeon et al. [14] presented a computer-aided diagnosis system for pigmented skin lesions. Dermacoscopic images were processed to obtain the lesion boundary and to quantify the degree of symmetry of the lesion. To facilitate the subsequent segmentation and feature quantification steps, they introduced a novel scheme to remove hair from the images, using luminance component in the LUV color space to differentiate hair from dark pigment. The segmentation technique was based on the clustering of a two-dimensional color space. A modified Fuzzy C-Means (FCM) technique called Orientation–Sensitive Fuzzy C-Means (OS-FCM) was used from the authors to perform the clustering. Symmetry of shape, color and texture were computed separately on a database consisting of 50 images of benign nevi and 50 images of malignant melanoma.

Ganster et al. [4] presented a system where as initial step the binary mask of the skin lesion was determined by several basic segmentation algorithms, combined together with a fusion strategy [4]. The algorithms used to segment the lesion are: global thresholding, dynamic thresholding, and a 3-D color clustering concept [4]. A set of features was then calculated to describe the malignancy of a lesion: global features (size and shape descriptors), color features and local features [4]. Significant features were selected from this set by application of statistical methods for feature selection [4]. The classification experiments were performed with a 24-NN classifier based on the derived features [4]. A notable characteristic of this work is the large dimension of the database. They had at their disposal overall 5363 skin lesion images, categorized into three classes. The three classes are: clearly benign lesions, dysplastic lesions and malignant lesions [4]. The training set for the classifier was a set of 270 lesions (90 images for each class). The test set was the entire database of 5363 lesions in three categories [4]. They obtained a mean recognition rate of 61%. To the best of our knowledge, this is the largest existing database on skin lesions, and these results constitute the state of the art in the field. This is the database on which we ran our experiments, and the results with which we compare our performance.

3 Spin Glass-Markov Random Fields

This section introduces briefly a probabilistic kernel method, Spin Glass-Markov Random Fields (SG-MRF) which was presented some years ago [8] and has shown promising results on biomedical applications [9]. We will first review the probabilistic approach to object recognition, then we will briefly review MRF and then we will describe the SG-MRF model. The interested reader can find more details in [8].

Probabilistic object recognition The probabilistic approach to appearance-based object recognition considers the image views $\boldsymbol{x}$ of a given object $\Omega_k, k = 1, \ldots \mathcal{K}$ as random vectors. Thus, given the set of data samples $\omega_k = \{\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_{n_k}^k\}$ and assuming they are a sufficient statistic for the pattern class Ω_k, the goal will be to estimate the probability distribution $P_{\Omega_k}(\boldsymbol{x})$ that has generated them. Then, given a test image $\boldsymbol{x}$, the decision step will be achieved using a Maximum A Posteriori (MAP) classifier $k^* = \mathrm{argmax}_{k=1}^{\mathcal{K}} P_{\Omega_k}(\boldsymbol{x}) = \mathrm{argmax}_{k=1}^{\mathcal{K}} P(\Omega_k|\boldsymbol{x})$, and, using Bayes rule,

$$k^* = \underset{k=1}{\overset{\mathcal{K}}{\mathrm{argmax}}}\, P(\boldsymbol{x}|\Omega_k)P(\Omega_k). \tag{1}$$

where $P(\boldsymbol{x}|\Omega_k)$ are the Likelihood Functions (LFs) and $P(\Omega_k)$ are the prior probabilities of the classes. In the rest of the paper we will assume that the priors $P(\Omega_k)$ are constant and the same for all object classes; thus the Bayes classifier (1) simplifies to

$$k^* = \underset{k=1}{\overset{\mathcal{K}}{\mathrm{argmax}}}\, P(\boldsymbol{x}|\Omega_k). \tag{2}$$

Probabilistic methods are philosophically optimal in the sense that with a posterior probability distribution over classes, selecting a maximum probability class will minimize the probability of error. A major problem in these approaches is that the functional form of the probability distribution of an object class Ω_k is not known a priori. Assumptions have to be made regarding the parametric form of the probability distribution, and parameters have to be learned in order to tailor the chosen parametric form to the pattern class represented by the data ω_k.

Markov Random Fields A possible strategy for modeling the parametric form of the probability function is to use Gibbs distributions within a Markov Random Field framework (MRF, [15]). MRF provides a probabilistic foundation for modeling spatial interactions on lattice systems or, more specifically, on interacting features. It considers each element of the random vector $\boldsymbol{x}$ (that in MRF terminology is called a *configuration*) as the result of a labeling of all the sites representing $\boldsymbol{x}$, with respect to a given label set:

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{x})\right), \qquad Z = \sum_{\{\boldsymbol{x}\}}\exp\left(-E(\boldsymbol{x})\right). \tag{3}$$

The normalizing constant Z is called the partition function, and $E(\boldsymbol{x}) = \sum_i f_i(x_i|x_{N_i})$ is the *energy function*. $P(\boldsymbol{x})$ measures the probability of the occurrence of a particular configuration $\boldsymbol{x}$; the more probable configurations are those with lower energies. Using MRF modeling for appearance-based object recognition, eq. (2) will become

$$k^* = \operatorname*{argmax}_{k=1}^{\mathcal{K}} P(\boldsymbol{x}|\Omega_k) = \operatorname*{argmin}_{k=1}^{\mathcal{K}} E(\boldsymbol{x}|\Omega_k). \tag{4}$$

Only a few MRF approaches have been proposed for high level vision problems such as object recognition [15, 8], due to the modeling problem for MRF on irregular sites (for a detailed discussion on this point, we refer the reader to [8]). SG-MRFs overcome this limitation and can be effectively used for appearance-based object recognition [8].

Spin Glass-Markov Random Fields SG-MRFs are a new class of MRFs that connect SG-like energy functions (mainly the Hopfield one [16]) with Gibbs distributions via a nonlinear kernel mapping. The resulting model overcomes many difficulties related to the design of fully connected MRFs, and enables us to use the power of kernels in a probabilistic framework. Consider $\mathcal{K}$ different object classes $\Omega_1, \Omega_2, \ldots, \Omega_{\mathcal{K}}$, and for each class a set of data samples $\omega_k = \{\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_{n_k}^k\}, k = 1, \ldots \mathcal{K}$. The SG-MRF probability distribution is given by

$$P(\boldsymbol{x}|\Omega_k) = \frac{1}{Z}\exp\left(\frac{1}{N}\sum_{\mu=1}^{p_k}\left[K_{d-G}(\boldsymbol{x}, \tilde{\boldsymbol{x}}^{(\mu)})\right]^2\right) \tag{5}$$

with K_{d-G} generalized Gaussian kernel

$$K_{d-G} = \exp\{-\rho d_{a,b}(\boldsymbol{x}, \boldsymbol{y})\}, \qquad d_{a,b} = \sum_{i=1}^{m}|x_i^a - y_i^a|^b \tag{6}$$

and prototypes given by the naive ansatz:

$$\{\tilde{\boldsymbol{x}}^\mu\}_{\mu=1}^{p_k \equiv n_k} = \{\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_{n_k}^k\}, \qquad \rho >> \Delta_{min}. \tag{7}$$

Note that SG-MRFs can be defined on features and on raw pixel data. The sites are fully connected, which ends in learning the neighborhood system from the training data instead of choosing it heuristically. Another key characteristic of the model is that in SG-MRF the functional form of the energy is given by construction. The interested reader will find the theoretical derivation of the model in [8].

4 Support Vector Machines

Support Vector Machines are state-of-the-art large margin classifiers which have gained popularity within visual pattern recognition. Here we provide a brief review of the theory behind this type of algorithm. For a more detailed treatment, we refer to [5].

Linear SVM Consider the problem of separating a set of training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, where $\boldsymbol{x}_i \in \Re^N$ 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 (the one with the lowest bound on the expected generalization error) 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 the following constrained minimization problem:

$$\underset{\boldsymbol{w},b}{\text{minimise}} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1, \forall i = 1, \dots m \tag{8}$$

Introducing Lagrange multipliers $\alpha_i (i = 1, \dots m)$ results in a classification function

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}\cdot\boldsymbol{x} + b\right). \tag{9}$$

where α_i and b are found by Sequential Minimal Optimization (SMO, [5]). Most of the α_i's take the value of zero; those $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In cases where the two classes are non-separable, Lagrange multipliers are introduced, $0 \leq \alpha_i \leq C, i = 1, \dots m$, where C determines the trade-off between margin maximization and training error minimization. It is possible also to give different costs to false-positive and false-negative errors, introducing the parameters C^+ and C^- respectively, instead of C [5].

Non-linear SVM To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a high dimensional feature space $\mathcal{H}$ by $\boldsymbol{x} \to \Phi(\boldsymbol{x}) \in \mathcal{H}$, 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 $K(\boldsymbol{x},\boldsymbol{y}) = \Phi(\boldsymbol{x})\cdot\Phi(\boldsymbol{y})$, a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{w}\cdot\boldsymbol{x}$ by the kernel function $K(\boldsymbol{x},\boldsymbol{y})$ in eqn. (9). This corresponds to constructing an optimal separating hyperplane in the feature space. The kernel $K(\boldsymbol{x},\boldsymbol{y})$ can be seen as a non-linear generalization of the Euclidean scalar product. Thus, choosing a kernel type corresponds to the choice of a similarity function for the classifier.

Kernel functions The impact of the choice of a kernel type on SVMs' performance has been clear since the introduction of the kernel trick for the non-linearization of the algorithm. Between the kernel types which were proposed at first, the *Gaussian Radial Basis Function (RBF) kernel*

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\{-\gamma||\boldsymbol{x}-\boldsymbol{y}||^2\}; \tag{10}$$

and the *polynomial kernel*

$$K(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}\cdot\boldsymbol{y} + c)^d; \tag{11}$$

has been widely used in computer vision. Since SVMs have started to be used for visual recognition, several researchers have proposed new kernel types and have studied their performances. In [17], Chapelle et al proposed two new types of exponential kernels: the *generalized Gaussian RBF kernel*

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\{-\gamma||\boldsymbol{x}^a - \boldsymbol{y}^a||^b\}, a \in \Re+, 0 < b \leq 2 \tag{12}$$

and the χ^2 *kernel*

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\{-\gamma\chi^2(\boldsymbol{x},\boldsymbol{y})\}, \qquad \chi^2 = \sum_i \frac{||x_i - y_i||^2}{||x_i + y_i||}. \tag{13}$$

These are the kernels we will use in this paper.

Multi-class SVM The extension of SVM from 2-class to M-class problems can be achieved following two basic strategies: In a *one-vs-others* approach, M SVMs are trained, each separating a single class from all remaining classes. In the second strategy, the *pairwise approach*, $M(M-1)/2$ two-class machines are trained. The pairwise classifiers are arranged in trees, where each tree node represents an SVM. Decisions can be made using a bottom-up tree similar to the elimination tree used in tennis tournaments [5]).

5 Experiments

In this section we present experiments that show the effectiveness of kernel methods for melanoma recognition. In the rest of the section we describe the database used (section 5.1), the experimental setup (section 5.2) and our experimental findings (sections 5.3- 5.5).

5.1 Database

We performed our experiments on the database created by the Department of Dermatology of the Vienna General Hospital [4]. 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).[2] The lesions of the classes 2 and 3 were all surgically excised and the ground truth was generated by means of histological diagnosis [4]. In order to have statistically significant results, we ran experiments with five different partitions, then we calculated the mean and standard deviation of the obtained recognition rates. This procedure has been adopted for all the experiments reported here. Figure 1 shows some exemplar images for each class.

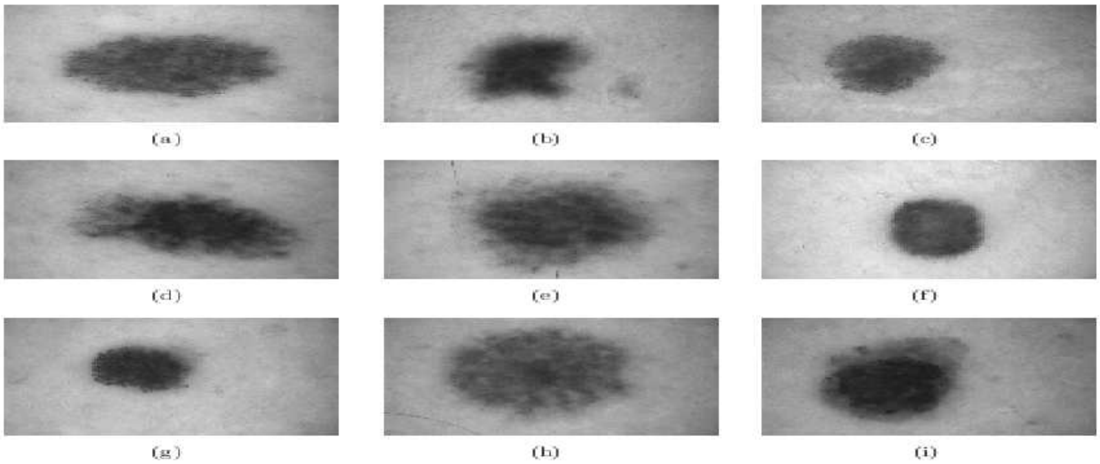

Figure 1: Examples of skin lesion's images used: (a,b,c) images of benign lesions, (d,e,f) images of dysplastic lesions,(g,h,i) images of malignant lesion. Note the high variability within one class, and the low variability between different classes. This makes the classification problem very challenging.

5.2 Experimental Setup

The three key components for an automated melanoma recognition algorithm are: segmentation/preprocessing, features extraction and classification. We describe below the general approach followed in this paper for each of these steps:

Segmentation/preprocessing: Following the approach proposed in [4], we didn't implement any preprocessing step such as color normalization or hair removal. As for the segmentation procedure, we used two different methods. The first consists in 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

[2]These numbers are not perfectly coincident with those reported in [4], where the database is said to be of 5363 images, but this difference should not affect the comparison between the two algorithms.

method is the one developed by Ganster et al. [4]. It consists of a binary mask determined by several segmentation algorithms combined together with a fusion strategy. We call the resulting images "mask-segmented". An example of the images obtained by these two segmentation techniques is in figure 2. Running experiments on these two types of images allows us to explore how the classification performance is affected by the quality of the segmentation process.

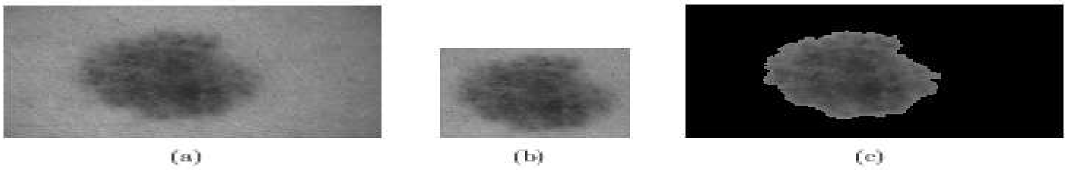

Figure 2: Examples of segmentation: (a) shows an example of an entire image, (b) the same image hand-segmented, (c) the same image mask-segmented

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 [10] and is defined as:

$$h_{R,G,B}(r,g,b) = mProb[R=r, G=g, B=b], \tag{14}$$

where R,G and B are the three color channels and m is the number of pixels in the image. The CH was computed by discretizing the colors within the image and counting the number of pixels for each color. For each series of experiments we used hue, rg, RG, RB and GB color histograms. Also, the resolution of the bin axes was varied for each representation consisting of 8, 16, 32, 64 (for bidimensional histograms we chose the resolution of each axis with the same bin value). In the following sections, we will always report the best result obtained.

The main idea of MFH is to calculate multidimensional histograms of the response of a vector of receptive fields [11]. 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 representation: the first consisted in Gaussian derivatives along x and y directions [11]:

$$G_x^\sigma = -\frac{x}{\sigma^2} G^\sigma(x,y), \qquad G_y^\sigma = -\frac{y}{\sigma^2} G^\sigma(x,y) \tag{15}$$

where $G^\sigma(x,y)$ is the Gaussian distribution [11] and with $\sigma = 1.0$ $(D_x D_y)$; the second consisted in Laplacian Gaussian operator [11]:

$$Lap(x,y) = (\frac{x^2}{\sigma^4} + \frac{1}{\sigma^2})G^\sigma(x,y) + (\frac{y^2}{\sigma^4} + \frac{1}{\sigma^2})G^\sigma(x,y) \tag{16}$$

where $G^\sigma(x,y)$ is the Gaussian distribution [11] and with $\sigma_1 = 1.0, 1.5, 3.0$, and $\sigma_2 = 2.0, 3.0, 6.0$ respectively $(Lp2\sigma)$. The bin axes' resolution was varied for each representation consisting of 8, 16, 32, 64 for Gaussian-filter MFH and 16, 32 for Laplacian-filter MFH.

Classification: We used SG-MRF and SVM algorithms (see section 3 and 4 respectively). For SG-MRF we learned the kernel parameters during the training stage using a leave-one-out strategy. For SVM we used the four kernel types described in section 4. The kernel parameters were chosen via cross validation.

We performed three series of experiments: in the first series, all the experiments were performed respecting the procedure reported by Ganster et al. [4]. The training set consisted of 270 images (90 for each class); the test set consisted of the whole database [4]. Note that training and test set are not disjoint; once again we underline that this follows the procedure proposed in [4], allowing for benchmarking. The results of these experiments are reported in section 5.3. For a fair evaluation of the algorithms, it is necessary to disjoin the training set from the test set. We performed therefore a second series of experiments using the following partition: the training set consisted of 270 images (90 for each class); the test set consisted of the remaining database. The results of these experiments are reported in section 5.4. Finally, a third series of experiments was performed for binary classification: if the aim of a CAD for skin lesions classification is to prescribe or not to prescribe the surgical excision of the lesion, it is reasonable to group dysplastic and malignant lesions into a common class, so to evaluate the performance of SVM with respect to this problem. In this series of experiments the database was then composed of two classes: class 1, coincident with class 1 of the previous experiments, and class 2, the union of the images of the class 2 and 3 of the previous experiments. The training set consisted of 180 images (90 for each class); the test set consisted of the remaining database. The results of these experiments are reported in section 5.5.

5.3 First series of experiments

We ran experiments using CH and MFH representation as features. The obtained recognition rates for hand-segmented and mask-segmented images using SG-MRF and SVM, with both features types are reported in table 1. Results for each class are averaged on five partitions. We also report the average of the recognition rate obtained class by class ("Mean Class"), and the overall recognition rate ("Overall"). For sake of clarity we report the results obtained in [4] too; note that these results were obtained on a single run.

A first comment is that SVM obtains the best result with respect to Ganster's method and SG-MRF, for both feature types and for both segmentation strategies. The best result, in terms of overall recognition rate, 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 result obtained by using SG-MRF is of 49.5%, obtained using mask-segmented images and MFH features; finally, the best performance obtained by using the Ganster's method is of 58%. These results clearly show the effectiveness of SVMs for melanoma recognition. A second comment is that SVM' performance varies considerably, depending on the kernel type used. For instance, using color features and hand-segmented images, the overall recognition rate goes from a minimum of 59.0% for the Gaussian kernel to a maximum of 76.0% for chi-squared 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, for the overall recognition rate, the kernels which obtains the worst performances tend to have the highest standard deviations, while the kernel with the best performance has the smallest one. This illustrates the importance of doing kernel selection in the training phase; the low standard deviation of the SVM's best results also shows the stability of our findings. We observe that the polynomial kernel obtains results comparable to 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 experiments. By comparing the hand-segmented overall 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. A final remark should be made on the poor performance of SG-MRF. This might be due to the dimension of the training set for each class; it could be possible that the probabilistic method needs a higher statistic in order to estimate properly the energy function. On the basis of these results we decided to not use this method in the remaining series of experiments.

Table 2 reports the confusion matrices for the best results obtained by each possible combination of (segmentation mask, feature type) and SVMs, plus the confusion matrix obtain by Ganster and that

			Class 1	**Class 2**	**Class 3**	**Mean Class**	**Overall**
Ganster et al. [4] (%)			59	53	73	61	58
CH features							
hand	**SG-MRF (%)**		43.2 ± 4.5	41.2 ± 2.1	95.1 ± 1.6	59.8 ± 17.6	46.1 ± 5.6
	SVM (%)	**poly**	91.7 ± 4.9	9.8 ± 7.3	5.5 ± 0.5	35.7 ± 28.0	74.9 ± 2.8
		gauss	65.7 ± 17.1	31.6 ± 16.0	49.5 ± 26.0	48.9 ± 9.8	59.0 ± 10.3
		gengauss	89.8 ± 20.4	15.6 ± 13.6	82.6 ± 14.6	62.7 ± 23.6	75.9 ± 14.0
		chi	90.0 ± 20.2	15.0 ± 12.3	89.1 ± 0.0	64.7 ± 24.9	**76.0 ± 13.7**
mask	**SG-MRF (%)**		48.6 ± 4.2	38.8 ± 3.4	94.1 ± 3.4	60.5 ± 17.0	47.7 ± 2.9
	SVM (%)	**poly**	80.1 ± 13.0	15.7 ± 13.7	29.5 ± 20.4	41.8 ± 19.6	67.1 ± 7.8
		gauss	71.9 ± 11.1	24.8 ± 12.7	45.0 ± 28.5	47.2 ± 13.6	62.6 ± 6.2
		gengauss	96.2 ± 4.0	11.0 ± 1.8	89.5 ± 0.9	65.6 ± 27.4	**80.2 ± 2.8**
		chi	68.6 ± 17.7	22.4 ± 7.5	62.6 ± 19.7	51.2 ± 14.5	59.9 ± 12.9
MFH features							
hand	**SG-MRF (%)**		39.2 ± 4.1	42.2 ± 3.1	94.5 ± 2.9	58.6 ± 18.0	40.8 ± 2.8
	SVM (%)	**poly**	85.3 ± 18.3	9.7 ± 8.5	19.8 ± 22.9	38.3 ± 23.7	66.9 ± 13.1
		gauss	55.7 ± 13.9	31.6 ± 17.1	54.1 ± 19.4	47.1 ± 7.8	51.1 ± 8.6
		gengauss	96.7 ± 2.8	11.7 ± 3.0	89.7 ± 0.9	66.0 ± 27.2	**80.7 ± 1.7**
		chi	80.8 ± 2.3	23.1 ± 4.0	93.1 ± 1.4	65.7 ± 21.6	70.3 ± 1.5
mask	**SG-MRF (%)**		49.3 ± 5.1	45.4 ± 4.0	94.5 ± 1.8	63.1 ± 15.7	49.5 ± 3.9
	SVM (%)	**poly**	80.5 ± 4.2	28.5 ± 14.9	22.0 ± 19.1	43.7 ± 18.5	69.7 ± 3.8
		gauss	80.9 ± 3.6	27.2 ± 13.5	25.3 ± 23.0	44.5 ± 18.2	69.8 ± 3.7
		gengauss	99.4 ± 0.1	9.6 ± 0.4	89.3 ± 0.4	66.1 ± 28.4	**82.5 ± 0.1**
		chi	96.7 ± 0.4	13.0 ± 1.6	90.5 ± 0.5	66.7 ± 26.9	81.0 ± 0.2

Table 1: Recognition results for the first series of experiments

Ganster et al [4]			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	2500	1347	410
class 2	324	531	155
class 3	14	12	70

Clinicians			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	4161	94	9
class 2	42	960	8
class 3	6	19	78

CH hand			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	3850.6	259.4	167.0
class 2	798.2	150.4	53.4
class 3	9.8	1.2	90.0

CH mask			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	4112.6	112.6	50.8
class 2	874.8	110.0	17.2
class 3	10.4	0.2	90.4

MFH hand			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	4184.8	45.5	45.8
class 2	861.6	116.8	23.6
class 3	9.8	0.6	90.6

MFH mask			
	Assigned		
True	**class 1**	**class 2**	**class 3**
class 1	4251.8	4.2	20.0
class 2	901.0	95.8	5.2
class 3	10.4	0.4	90.2

Table 2: Confusion matrices for the first series of experiments

relative to clinicians' performance on the database [4].[3] For both segmentation techniques and feature types, we see that SVM outperforms Ganster's method for class 1 and class 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, while 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 in order to recognize correctly dysplastic lesions, while both feature types seem to be effective for separating benign and malignant lesions.

5.4 Second series of experiments

Experiments reported in section 5.3 were performed on a not-disjoint experimental set. This was done in order to compare fairly with the results reported in [4], but this strategy doesn't allow to evaluate properly the generalization capability of our 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. We used SVM as classification algorithm, with kernel functions chi-squared, the Gaussian and the generalized Gaussian ones.

The classification results for hand-segmented and mask-segmented images using SVM, with both features types, are showed in table 3. We also report the average of the recognition rate obtained class by class ("Mean Class"), and the overall recognition rate ("Overall").

We see that the best overall recognition rate is of $82.9 \pm 0.9\%$. This result must be compared with the best overall recognition rate obtained on training and test set not-disjoint (section 5.3, table 1), which was of $82.5{\pm}0.1\%$. 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 before, 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.

An important observation is that for the overall, class 1 and class 2 recognition rates (and standard deviations) the values in table 2 are statistically equivalent with the values obtained using training and test set not-disjoint (see table 1). This doesn't hold for class 3, where the high standard deviation does not permit a comparison within the two series of experiments. We believe this is due to the low statistic of the database for the class 3. Indeed, class 3 consists of 101 lesions, as to say one order of magnitude less than dysplastic or benign lesions. Particularly, in these experiments we used 90 images for training and 11 for test. This leads to a high variability on the class 3 recognition rate, and consequently to a very high standard deviation.

5.5 Two class experiments and ROC analysis

As reported in [4], during routine clinical practice in the Vienna General Hospital, the lesion is not surgically excised if is achieved by a consensus of three experienced dermatologists that the lesion's diagnosis is "benign" [4]. The lesions named "dyslastic" are still considered as benign, but are so-called precursors to malignant melanoma. Since this category represents skin lesions with an increased risk to turn into a melanoma, the category receives its own class label in the clinical diagnosis. The lesions classified as dysplastic are all surgically excised, as it is done for malignant ones [4].

Given that the purpose of a CAD system is to prescribe or not the surgery, it could be useful 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 lesions images. The CAD will suggest the surgical

[3]For more details on the number of images used in the these last two confusion matrices we refer the reader to [4].

			Class 1	Class 2	Class 3	Mean Class	Overall
CH features							
hand	**SVM** (%)	**gauss**	65.3 ± 17.4	29.1 ± 15.8	45.5 ± 21.3	46.63 ± 10.47	58.8 ± 11.4
		gengauss	89.8 ± 20.6	8.2 ± 16.9	5.45 ± 12.2	34.5 ± 27.7	**75.0 ± 13.8**
		chi	89.8 ± 20.5	6.8 ± 13.9	7.3 ± 16.3	34.6 ± 27.6	74.8 ± 14.3
mask	**SVM** (%)	**gauss**	69.0 ± 14.5	27.3 ± 17.0	32.8 ± 22.8	43.0 ± 13.1	62.8 ± 7.1
		gengauss	94.5 ± 5.2	2.9 ± 2.5	9.1 ± 12.9	35.5 ± 29.6	**79.1 ± 3.0**
		chi	68.4 ± 17.7	19.0 ± 10.6	43.6 ± 33.7	43.7 ± 14.3	60.2 ± 12.3
MFH features							
hand	**SVM** (%)	**gauss**	54.8 ± 15.0	33.8 ± 21.0	43.6 ± 17.5	44.1 ± 6.1	51.0 ± 9.26
		gengauss	96.5 ± 2.9	2.6 ± 3.3	5.5 ± 8.1	34.9 ± 30.8	**79.5 ± 1.8**
		chi	80.4 ± 2.4	15.5 ± 4.3	36.4 ± 12.8	44.1 ± 19.1	68.8 ± 1.6
mask	**SVM** (%)	**gauss**	81.3 ± 2.8	25.0 ± 13.0	30.9 ± 41.0	45.7 ± 17.9	70.8 ± 4.0
		gengauss	99.4 ± 0.1	8.5 ± 4.4	80.0 ± 44.7	62.6 ± 27.6	**82.9 ± 0.9**
		chi	92.5 ± 1.4	16.8 ± 2.0	81.8 ± 40.7	63.7 ± 23.7	78.8 ± 1.5

Table 3: Recognition results for the second series of experiments

excision for the lesions in this class.
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); the test set consisted of the remaining database. For SVM we used the chi-square, Gaussian and generalized Gaussian kernel types, described in section 4. The kernel parameters were chosen via cross validation; the obtained results were analized using the Receiver Operating Characteristic (ROC) analysis [18].

The ROC analysis is a widely used method in the medical community, for estimating the accuracy of a binary decision-making process. A ROC curve is a plot of the True Positive Fraction (TPF) versus its False Positive Fraction (FPF) [18]; for a given classifier, it is obtained varying the threshold associated with its decision function and the Area Under the ROC Curve (AUC) is a summary measure of overall diagnostic performance. The AUC can take values from 0.5 (no apparent accuracy) to 1.0 (perfect accuracy) [19]. In this case the TPF are the correctly classified images of class 2 and the FPF are the incorrectly classified images of class 1. In our experiments we posed the cost parameter (C parameter, section 4) equal to 1 and we varied the C^+/C^- ratio (see section 4) from 1 to 9 in order 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. Table 4 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 5 reportes 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 chi-square 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 the 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

Kernel	Sensitivity - CH (%)	Specificity - CH (%)
chi	99.78 ± 0.11	76.46 ± 7.37
gauss	64.76 ± 19.69	85.70 ± 3.80
gengauss	100.00 ± 0.00	99.23 ± 0.25
Kernel	**Sensitivity - MFH (%)**	**Specificity - MFH (%)**
chi	77.87 ± 8.59	97.97 ± 0.75
gauss	85.88 ± 1.35	57.42 ± 9.51
gengauss	100.00 ± 0.00	99.47 ± 0.35

Table 4: Sensitivity and specificity results for the third series of experiments

Kernel	AUC for CH features
chi	0.882 ± 0.037
gauss	0.832 ± 0.039
gengauss	0.995 ± 0.003
Kernel	**AUC for MFH features**
chi	0.987 ± 0.005
gauss	0.741 ± 0.052
gengauss	0.997 ± 0.002

Table 5: AUC average values for the third series of experiments

99.47 for MFH.
We can conclude this section saying that all our experimental findings clearly indicate the excellent performances of SVM for melanoma detection.

6 Conclusions

In this paper we presented a learning approach to melanoma recognition. To this purpose, we proposed two kernel-based classification algorithms: a probabilistic one, spin glass-Markov random fields, and a discriminative one, support vector machines. Both methods have proved successful on visual recognition problems like object recognition. The two classifiers were tested on a database of more than 5000 images, using two feature types and two segmentation methods. Our results show that SVM obtains an improvement in recognition rate of more than 20% compared to what reported in [4], which to our knowledge constitutes the state of the art in the field. Moreover, on two classes out of three, SVM achieves recognition results comparable to those obtained by skilled clinicians.
A second series of experiments, with disjoint training and test set, confirmed the effectiveness of SVM. Moreover, a series of two class experiments showed that SVM gives surprising results in sensitivity and specificity, indicating that SVM could be an excellent aid for the physicians in the dichotomic choice of prescribe or not the surgical excision of the lesion.

In the future we plan to conduct similar experiments using shape descriptors, and finally to experiment with cue integration schemes, in order to test the effectiveness of different types of information and eventually to reproduce the ABCD method, which is followed by the dermatologists in every day clinical practice.

References

[1] Informations available at the World Health Organization website: http://www.who.int

[2] Burroni M, Corona R, Dell'Eva G, Sera F, Bono R, Puddu P, et al. Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study. Clinical Cancer Research Vol. 10, 1881-1886, March 2004.

[3] Rigel DS, Carucci JA. Malignant Melanoma: Prevention, Early Detection, and Treatment in the 21st Century. CA Cancer J Clin 2000; 50:215-236.

[4] Ganster H, Pinz A, Roddothrer R, Wildling E, Binder M, Kittler H. Automated Melanoma Recognition. IEEE Trans on MI, 20, 3, march 2001.

[5] Vapnik V. Statistical learning theory. Wiley and Son, 1998.

[6] Scholkopf B, Smola AJ. Learning with kernels. 2001, the MIT Press.

[7] Wallraven C, Caputo B, Graf A. Recognition with Local features: the kernel recipe. Proc. ICCV03.

[8] Caputo B. A new kernel method for object recognition: spin glass Markov random fields. PhD thesis, Stockholm, November 2004. Available at http://www.nada.kth.se/∼caputo

[9] Caputo B, La Torre E, Bouattour S, Gigante GE. A New Kernel Method for Microcalcification Detection: Spin Glass- Markov Random Fields. Proc. of MIE02, Budapest, August 2002.

[10] Swain M, Ballard D. Color Indexing. Internal Journal of Computer Vision, 7(1), 1991, pp 11-32.

[11] Schiele B, Crowley JL. Recognition without correspondence using Multidimensional Receptive Field Hisograms. IJCV, 36(1), 2000, pp 31-52.

[12] Grana C, Pellacani G, Cucchiara R, Seidenari S. A New Algorithm for Border Description of Polarized Light Surface Microscopic Images of Pigmented Skin Lesions. IEEE Trans on MI, 22, 8, August 2003.

[13] Maglogiannis I, Pavlopoulos S, Koutsouri D. An integrated computer supported acquisition, handling, and characterization system for pigmented skin lesions in dermatological images. IEEE Trans Inf Technol Biomed. 2005 Mar;9(1):86-98.

[14] Schmid-Saugeon P, Guillod J, Thiran JP. Towards a Computer-Aided Diagnosis System for Pigmented Skin Lesions. Computerized Medical Imaging and Graphics, 27(1):65-78, January 2003.

[15] Li SZ. Markov random field modeling in computer vision. Springer-Verlag, 1995.

[16] Amit DJ. Modeling Brain Function. Cambridge University Press, Cambridge, USA, 1989.

[17] Chapelle O, Haffner P, Vapnik V. SVMs for histogram based image classification. *IEEE Transaction on Neural Networks*, 9, 1999.

[18] Van Erkel AR, Pattynama PMT. Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. European J Radiol 1998; 27: 88-94.

[19] Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 1982: 143: 29-36.