Bag-of-Features Classification Model for the Diagnose of Melanoma in Dermoscopy Images Using Color and Texture Descriptors

Catarina Barata[1], Jorge S. Marques[1], and Teresa Mendonça[2]

[1] Institute for Systems and Robotics, Instituto Superio Técnico, Portugal
ana.c.fidalgo.barata@ist.utl.pt, jsm@isr.ist.utl.pt
[2] Faculdade de Ciências, Universidade do Porto, Portugal
tmendo@fc.up.pt

Abstract. Melanoma detection using medical oriented approaches has been a trend in skin cancer research. This paper uses a Bag-of-Feature model for the detection of melanomas in dermoscopy images and aims at identifying the role of different local texture and color descriptors. This is a medical oriented approach and the reported results are promising (Sensitivity = 93%, Specificity=85%), showing the ability of this method to describe medical dermoscopic features. Moreover, the results show that color descriptors outperform texture ones.

Keywords: Melanoma, Dermoscopy, Bag-of-Features, Feature Extraction, Medical Image Analysis.

1 Introduction

Dermoscopy is a widely accepted diagnostic technique used by dermatologists to help them improve the early diagnosis of melanomas. This non-invasive microscopy approach can be performed using different inspection methods. One of them, a digital imaging system, can be used to acquire magnified images of the skin lesions [1]. Over the last two decades, several Computer-Aided Diagnosis (CAD) systems, that use these images as an input, have been proposed to help dermatologists distinguish between benign and malignant skin lesions. These systems share a set of steps: artifact removal, lesion segmentation, feature extraction, feature selection and classification, being the last three steps the basis of a pattern recognition method [2].

More recently, a different trend of dermoscopy CAD systems has emerged. These systems aim to reduce the gap between the medical and engineering knowledge, by trying to mimic the dermatologists behavior when diagnosing a skin lesion. Some of the developed systems try to detect specific dermoscopic cues (called local dermoscopic features [1]) such as specific coloration (e.g, blue-white veil [3]) or differential structures (e.g., pigment network [4]). These local dermoscopic features are considered the letters of the dermoscopic alphabet and they are the criteria assessed in clinical algorithms (ABCD rule [5] and 7-point

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 547–555, 2013.
© Springer-Verlag Berlin Heidelberg 2013

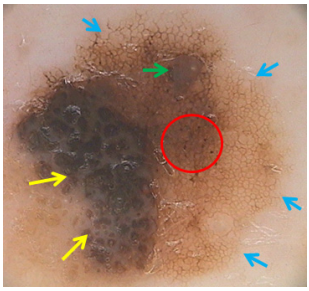

Fig. 1. Dermoscopic features: dots (red circle); pigment network (blue arrows), cobblestone pattern (yellow arros) and homogeneous pattern (green arrow)

checklist [6]) to distinguish between melanomas and benign lesions. Alternatively, dermatologists can assess specific patterns (e.g, reticular, cobblestone, parallel and homogeneous) that are characteristic of certain pigmented skin lesions and allow a fast and simple categorization [1].This method is called pattern analysis and its reproduction has also been attempted with promising results in the recognition of different patterns [7]. Fig. 1 exemplifies some of the commonly detected local dermoscopic features and patterns.

Despite their promising results, as far as the authors know, only one system performs both the dermoscopic criteria detection and the lesion classification using the 7-point checklist classification method [8]. This evidence suggests that combining the information provided by the detected cues with the medical knowledge and algorithms in order to develop a lesion classification system is a challenging task. Bag-of-Features (BoF) [9] is an image analysis and classification method, that can be used to overcome this difficulty. In this method, a lesion is represented by a set of local features, each associated to a small region inside the lesion. Therefore, different dermoscopic cues can be identified and characterized independently (assuming spatial independence), allowing the integration of medical knowledge in the CAD system. Moreover, due to its properties, BoF is also a classification method that simplifies the development of a medical inspired classification system. This method has already been used successfully in the melanoma identification problem [10]. However, a comparison between different types of descriptors has not yet been performed.

This paper describes a BoF model for the classification of melanomas using two different types of local descriptors: texture and color. The performance of both descriptors is compared in order to assess their ability to describe the different dermoscopic features. The paper is organized as follows. Section 2 describes the BoF system as well as the different descriptors tested. Section 3 describes the experimental setup and presents the results obtained. Finally, Section 4 concludes the paper.

2 Classification System

This section introduces de BoF method and the several descriptors tested. Fig.2 shows an overview of the BoF method.

2.1 Bag-of-Features

BoF is a well known classification model used successfully in several challenging classification tasks such as scene recognition and object identification [9]. The main idea behind this strategy is that an image can be modeled by a set of local features. To extract this set of local features the image must be sampled into smaller regions (patches) and a descriptor is computed for each of the patches. Two different sampling strategies are commonly used: sparse and dense sampling. In this work the strategy used is dense sampling, which consists of extracting the square patches using a regular grid (see Fig. 2). After the sampling process, several features can be used to describe the patches. In the following section the color and texture descriptors used are discussed.

After extracting the patches and computing their feature vectors, a dermoscopy image I with N patches will be represented by a family of patch feature vectors. This representation is not practical because the number of patches varies between images, thus it is not possible to train a classifier. In order to be able to perform this task, a visual vocabulary has to be constructed. This is usually done using a clustering algorithm like K-means and the computed clusters are called *visual words*. The dictionary is then used to assign a specific *visual word* to each patch in the training set. By counting the occurrence of each *visual word* in a given image it is possible to describe it as a histogram of *visual words* frequency. This histogram will act as the feature vector of the image and it will be used to train the classification algorithm. The dictionary size is one of the key factors for BoF performance. On one hand, a large dictionary has a good discriminative power but is less generalizable and requires greater processing time. On the other hand, a smaller dictionary may be more generalizable but it is not so discriminant. Therefore, three dictionary sizes, i.e. three numbers of clusters ($K \in \{100, 200, 300\}$) are tested.

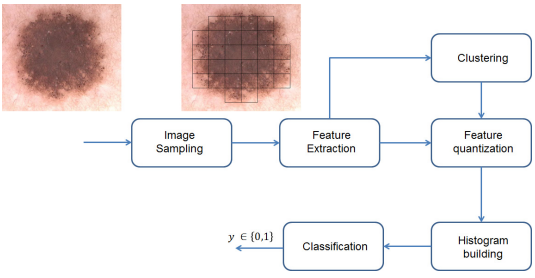

Fig. 2. BoF system overview

Several classification algorithms can be used in the decision step. In this work, the k-Nearest Neighbor (kNN) algorithm is employed. The optimal parameters for this classifier (the number of neighbor k and the feature comparison distance) are search in the interval $k \in \{5, 7, ..., 25\}$ and between three different distances {Euclidean, Kolmogorov, Kullback-Leibler}.

2.2 Patch Descriptors

The extracted patches must be characterized by appropriate descriptors that are able to represent the dermoscopic criteria assessed by dermatologists. Two different classes of descriptors can be used to represent them: color and texture descriptors.

Texture descriptors represent the spatial organization of intensity in an image, which is directly related with the identification of shapes and structures. Therefore, this class of descriptors is suitable for describing local dermoscopic structures such as pigment network, dots and streaks. There are several commonly used texture descriptors. In this work four of them are tested and their performances compared. The used descriptors are Gray Level Co-occurrence Matrix (GLCM) [11], Gabor filters [12], Laws masks (using the nine combinations of Level, Edge and Spot masks) [13] and gradient. For GLCM the most common statistics (entropy, energy, contrast, correlation and homogeneity) are computed and used as features. In the cases of Gabor filters and Laws masks the computed features are the mean and standard deviation. Finally, the computed features using the gradient information are its phase (h_ϕ) and amplitude (h_a) histograms. The features are extracted from gray-level images obtained by selecting the RGB channel with the highest entropy value [14]. Since the performance of the classification method greatly depends on the discriminative power of the feature vectors, the specific parameters of each descriptor are optimized. Table 1

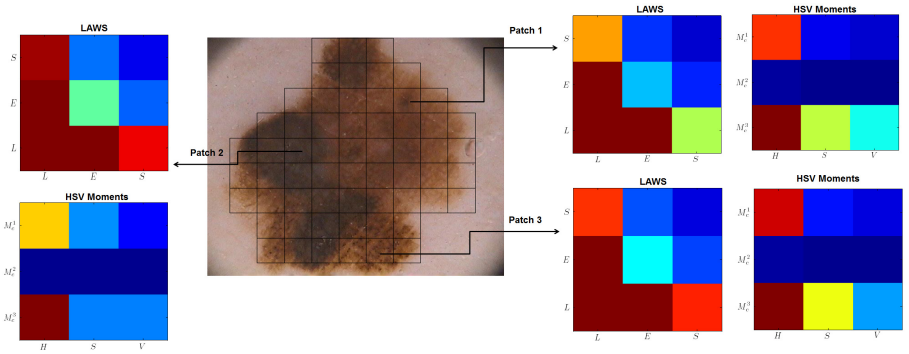

Fig. 3. Example of local descriptors for three different patches. The represented features are the three HSV channels color moments and the mean computed over the patches filtered using each of the nine combinations of Laws masks. The color scale in each mask ranges from dark blue (lowest value) to dark red (highest value).

Table 1. Texture descriptors and respective parameters

Descriptor	Parameters
GLCM	Gray levels: $G \in \{24, 34, ..., 64\}$
Gabor Filters	Number of scales: $s \in \{1, 2, ..., 5\}$ Number of orientations: $o \in \{2, 3, ..., 10\}$
Laws Masks	Kernel size: 3×3 or 5×5
Gradient	Number of bins: $B_{a,\phi} \in \{15, 25, 35, 45\}$

summarizes the descriptors, the tested parameters and their range for each of the texture descriptors.

Appropriate color descriptors can be used to characterize localized atypical coloration. The color analysis can be performed in different color spaces, each one with different properties. RGB is the most popular color space. However, it has a series of disadvantages: it is not perceptually uniform; it depends on the acquisition setup and shows correlation among the three color channels. There are some alternatives like HSV/I color spaces that perform a description of color similarly to human one or La^*b^*/L^*uv that are perceptually uniform color spaces. Another alternative is the Opponent (Opp) color space [15], which is inspired in the human visual process. Since it is not known which is the best color space for this specific problem all the previous ones are tested in this work.

Different color descriptors have been proposed for local analysis [15]. In this work two descriptors are employed: color histograms and moments. The color histogram feature vector results from the concatenation of the three color components 1-D histograms. As in the case of texture descriptors, the specific parameters of the descriptor are optimized. In the case of the color histograms, the parameter is the number of bins $B_c \in \{15, 25, 35, 45\}$. Color moments result from assuming that the color distribution in an image can be seen as a probability distribution, thus it can be characterized by a set of unique statistics. Three order color moments are used in this work: mean (M_c^1), standard deviation (M_c^2) and skewness (M_c^3). The previous moments are computed separately for each channel of the color spaces used.

Fig.3 shows an example of the Laws and HSV moments descriptors for three different patches. All the patches have different characteristics and it is clear that the exemplified features are different between the three of them.

3 Results

The BoF algorithm and the descriptors are tested on a dataset of 176 dermoscopy images (25 melanomas and 151 benign lesions). These images were acquired with a digital acquisition system and a magnification of 20×, during clinical exams performed on different patients at Hospital Pedro Hispano. Each image was classified by an experienced dermatologist and a ground truth label was created. To make the classification system independent from the segmentation all images were manually segmented. Different patch sizes are tested ($\delta \in \{20, 40, 60, 80\}$) and patches that are more than 50% outside the lesion are discarded.

The training and test process is performed using a 10-fold stratified cross-validation method. The dermoscopy images were evenly split between the 10 subsets, each one containing approximately the same number of melanomas and non-melanomas. The reported results are the average of 10 training-testing processes, each one using a different fold for testing and the remaining nine folds for training. To tackle the class unbalance problem, artificial melanoma examples were created by repeating the features associated with the melanomas and adding a small amount of noise.

The descriptors evaluation metrics are the Sensitivity (SE) and Specificity (SP). The best pair of results for each descriptor is selected using the following cost function

$$\mathcal{C} = \frac{c_{10}(1 - SE) + c_{01}(1 - SP)}{c_{10} + c_{01}} \quad , \tag{1}$$

where c_{10} is the cost of an incorrectly classified melanoma and c_{01} is the cost of an incorrectly classified non-melanoma. c_{10} is experimentally set to be $1.5c_{01}$ and $c_{01} = 1$. This function represents the trade-off between SE and SP and at the same time gives more weight to the correct classification of melanomas, since a false negative error is more grievous than a false positive one.

Fig.4(left) shows the best cost results obtained for each descriptor. It is observed that the best texture descriptors are Gabor (SE=98%, SP=64%) and Laws (SE=88%, SP=77%) and that their performance is similar to the one of most color moments. The best results are achieved with color histograms (h_{L^*uv}, SE=100%, SP=75% and $h_{La^*b^*}$, SE=93%, SP=85%) and most of them perform better than their corresponding color space moments (exception of the Opp. space). Fig.4(right) shows the classification results after combining pairs of descriptors. The descriptors were combined using an early fusion approach where two different feature vectors are concatenated. In this case, the combined feature vectors were the color histograms with their corresponding moments (identified in Fig.4(right) by the respective color spaces) and the two best texture features, Gabor and Laws (labeled as $Text$ in Fig.4). The performance of each separate descriptor is also represented in the graphic (red asterisks) and it is possible to notice that for most cases the fusion improves the results. In the cases where it does not happen, the cost is still better than the one obtained using only the worst descriptor of the pair.

Color and texture descriptors were also combined using a late-fusion strategy, i.e. the final decision is made by combining the output of separate classifiers. Different rules can be used to combine classifiers. In this paper the selected rule

Table 2. Late fusion $\mathcal{C}$ results of $Text$ with each of the color spaces features. The $\mathcal{C}$ scores for each of the fused classifiers are also shown.

		RGB	HSV	La^*b^*	HSI	L^*uv	Opp
		0.099	0.121	0.125	0.131	0.141	0.121
$Text$	0.146	0.099	0.124	0.122	0.128	0.129	0.112

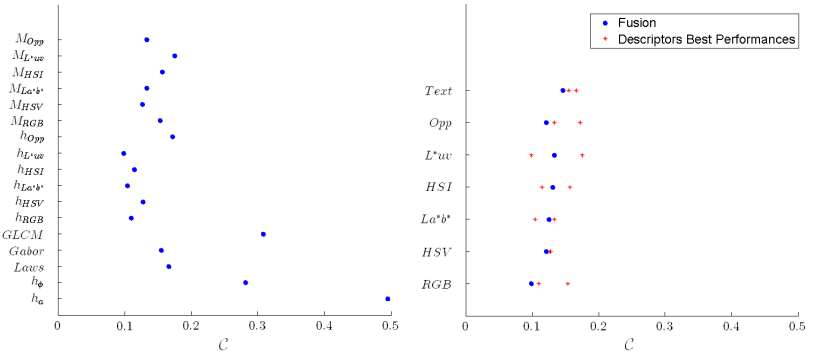

Fig. 4. Best cost results for single descriptors (left) and fusion of descriptors (right): Gabor+Laws, labeled as $Text$, and color moments with their respective color histograms, labeled with the color space. The best results obtained with each descriptor separately are also represented in the graphic.

was the Sum rule [16]. The pairs of descriptors that led to the results shown on Fig.4 (right) were combined and the fusion results can be seen on Table 2. For most cases, the fusion improved the results, which suggests that color and

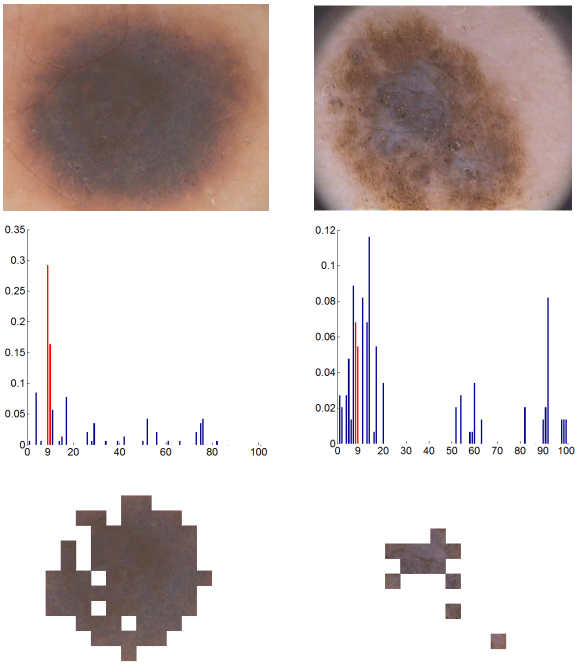

Fig. 5. Analysis of M_{Opp} dictionary obtained using 40×40 patches: melanomas (1st row); histograms of visual word frequencies (2nd row) and visual words (red bins of the histogram) corresponding patches (3rd row).

texture descriptors complement one another. The best results are achieved with $Text + RGB$ (SE=100%, SP=75%).

Although a direct comparison with state of the art methods is not possible due to different datasets, we can still assess if our results are within the same range of values. Situ et al. [10] use the BoF method to diagnose melanomas and achieve a SE=86% and SP=85% on a dataset of 1505 images (407 melanomas). Iyatomi et al. [2] achieved a SE=85.9% and a SP=86% on a dataset of 1258 images (198 melanomas), using a global approach. Finally, Di Leo et al [8] achieved a SE=83% and SP=76% on a dataset of 287 images (173) melanomas, with their automatic implementation of the 7-point checklist method [6]. Our results (SE=93% and SP=85%) are within the same range as the ones achieved with these three different approaches.

The different clusters (*visual words*) found during the dictionary construction can provide interesting information. Fig. 5 exemplifies a simple analysis of these words for two melanomas obtained using the system trained with the M_{Opp} descriptor. By extracting the patches associated with two random consecutive words (highlighted red bins in Fig.5 (2nd row)) it is possible to notice that they correspond to a well known dermoscopic feature: blue-whitish veil (see Fig.5 (3rd row)), which is one of the hallmarks of melanomas [1]. This evidence suggests that this method has the potential to be used as a dermoscopic feature detector.

4 Conclusions

This paper described a BoF model for the classification of melanocytic lesions. Several texture and color descriptors were evaluated separately and it was concluded that color histograms achieve better results (SE=93%, SP=85%). Descriptors fusions also achieved interesting results. A simple analysis of one of the dictionaries demonstrated that BoF can be used to identify dermoscopic criteria, suggesting that this approach can be seen as medical oriented one.

Future work should rely on testing sparse sampling methods and high-level descriptors as well as performing a deep analysis of *visual words* and use them to identify dermoscopic criteria.

Acknowledgments. This work was supported in the scope of the FCT grant SFRH/BD/84658/2012 and projects PTDC/SAUBEB/103471/2008 and PEst-OE/EEI/LA0009/2011.

References

1. Argenziano, G., et al.: Interactive atlas of dermoscopy (2000)
2. Iyatomi, H., et al.: An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. CMIG 32(7), 566–579 (2008)
3. Celebi, M., et al.: Automatic detection of blue-white veil and related structures in dermoscopy images. CMIG 32(8), 670–677 (2008)

4. Barata, C., et al.: A system for the detection of pigment network in dermoscopy images using directional filters. IEEE TBME 59(10), 2744–2754 (2012)
5. Stolz, W., et al.: ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Euro. J. Dermatology 4, 521–527 (1994)
6. Argenziano, G., et al.: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arc. Dermatology 134, 1563–1570 (1998)
7. Serrano, C., et al.: Pattern analysis of dermoscopic images based on markov random fields. PR 42, 1052–1057 (2009)
8. Di Leo, G., et al.: Automatic diagnosis of melanoma: A software system based on the 7-point check-list. In: Proc. 2010 43rd Hawaii ICSS, pp. 1818–1823 (2010)
9. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. 9th IEEE ICCV, pp. 1470–1477 (2003)
10. Situ, N., et al.: Evaluating sampling strategies of dermoscopic interest points. In: Proc. 8th ISBI, pp. 109–112 (2011)
11. Haralick, R.M., et al.: Textural features for image classification. IEEE TSMC 3, 610–621 (1973)
12. Arivazhagana, S., et al.: Texture classification using gabor wavelets based rotation invariant features. PR Letters 27, 1976–1982 (2006)
13. Laws, K.: Rapid texture identification. In: Proc. SPIE CIPMG (1980)
14. Silveira, M., et al.: Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. IEEE Journal. STPS 3, 35–45 (2009)
15. van de Sande, K., et al.: Evaluating color descriptors for object and scene recognition. IEEE TPAMI 32, 1582–1593 (2010)
16. Kittler, J., et al.: On combining classifiers. IEEE TPAMI 20, 226–239 (1998)