The Role of Keypoint Sampling on the Classification of Melanomas in Dermoscopy Images Using Bag-of-Features

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

[1] Institute for Systems and Robotics, Instituto Superio Técnico, Portugal
ana.c.fidalgo.barata@ist.utl.pt, jsm@isr.ist.utl.pt
[2] Hospital Pedro Hispano, Matosinhos, Portugal

Abstract. Integrating medical knowledge on a Computer Aided-Diagnosis systems for the detection of melanomas is an essential factor for the acceptance of the system by the medical community. Bag-of-Features, a popular classification method based on a local description of an image, can be used as a means to integrate medical knowledge while developing an automatic melanoma classification system. An important step of this algorithm is the correct identification of discriminative regions, due to the great impact that it has on the algorithm's performance. This paper aims at comparing different strategies for the extraction of interest regions. The achieved results show that texture-based detectors perform better than a dense sampling strategy, achieving Sensitivity= 98% and Specificity= 86%.

Keywords: Melanoma, Dermoscopy, Computer-Aided Diagnosis Systems, Bag-of-Features, Keypoints Detection.

1 Introduction

Melanoma is the deadliest form of skin cancer. Its great potential to rapidly metastasize and growing incidence rates make melanoma one of the 21st century diseases. Nowadays, the goal of dermatologist is to diagnose melanomas in their early stage, since it is less probable that it has already spread to other organs or tissues. One of the most popular techniques used by dermatologist to diagnose skin lesions is Dermoscopy. This is a microscopy technique that allows the visualization of different dermoscopic structures and pigmentations that would be otherwise invisible to the naked eye [1]. These dermoscopic features are, in most cases, discriminative of the type of skin lesion that is being analyzed and are the backbone of the medical algorithms proposed for the diagnose of dermoscopy images (e.g., ABCD rule [2] and 7-point checklist [3]).

A Computer Aided-Diagnosis (CAD) system for the detection of melanomas can benefit from using the medical knowledge associated with the dermoscopic features and their relative importance. There are two ways in which the medical knowledge can be incorporated in the diagnosis system. The first is based on

J.M. Sanches, L. Micó, and J.S. Cardoso (Eds.): IbPRIA 2013, LNCS 7887, pp. 715–723, 2013.

detecting the dermoscopic features, such as pigment network [4], blue-whitish veil [5] and globular pattern [6], and try to analyze them as a dermatologist do (assess their shape, color and distribution throughout the lesion). Then this information can be used to perform one of the medical algorithms cited above. However, detecting all the important dermoscopic features, classify them, combine all the information and finally apply the medical algorithm might not be easy to perform. The alternative approach follows a different direction and uses a popular image retrieval and object recognition method called Bag-of-Features (BoF) [7][8]. This method represents the lesion by a set of local descriptors, each of them associated to an interest region inside the lesion. Each interest regions can be interpreted as one of the dermoscopic cues used by dermatologist and will be separately characterized from the others. Assuming the former, it is then possible to consider that medical knowledge is being integrated in the CAD system. The benefit of using BoF against the previous strategy is that it is a classification algorithm whose parameters can be learned from the data. Therefore, it reduces the effort of developing a diagnose system.

One of the main steps of BoF is finding the interest regions. A simple way to detect these regions is to search for interest points (called sampling) and extract the regions around them. This strategy converts the previous problem into a keypoint finding one. Different strategies can be used to detect keypoints in an image. These strategies might influence the performance of BoF and its discriminative power, since it is desirable that the sampling strategy used is able to detect dermoscopic cues that characterize a lesion. The focus of this paper is to address the two common sampling strategies (sparse and dense sampling [9]) and determine which one performs better in the melanoma classification problem. The paper is organized as follows. Section 2 describes the BoF approach used and Section 3 explains the sampling approaches. The results and respective discussion are shown in Section 4 and Section 5 concludes the paper.

2 Bag-of-Features

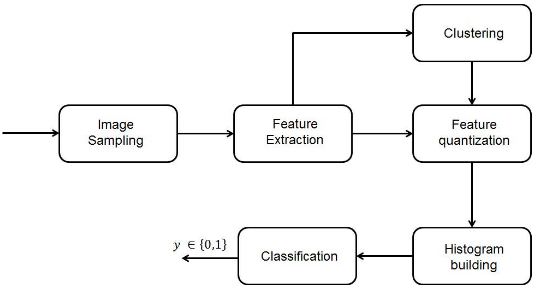


Fig. 1. BoF system overview

Fig.2 shows the block diagram of a BoF classification system [7]. The "Image Sampling" block of BoF is divided in two sequential tasks. The first task consists in detecting the keypoints. The comparison of possible sampling strategies is the focus of this paper and the ones used are described in the next section. After the detection of the keypoints, the next step is to extract the interest regions. These regions are square patches of size $\delta \times \delta$ centered on the detected keypoints. Patches which area is more than 50% outside the lesion are discarded.

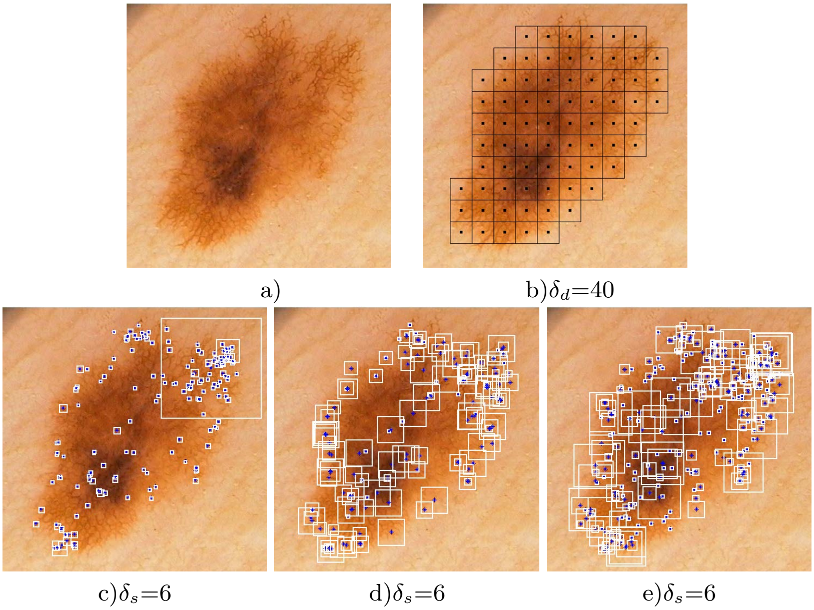
After the sampling process, each one of the detected regions is described by a set of features. In this paper two different types of color descriptors (color histograms [10] and color moments [11]) and six color spaces (RGB, HSV/I, La*b*, L*uv and Opponent [10]) are used. The color histograms (h_c) are a combination of three 1-D color histograms with a dimension B_c, one for each color component. B_c was optimized and searched in the intervals $B_c \in \{15, 25, 35, 45\}$ for dense sampling and $B_c \in \{5, 15, 25\}$ for sparse sampling. The color moments used (M_c) are the traditional first three order color moments (mean, standard deviation and skewness). These moments are computed for each of the three color components, leading to 9 moments per color space. All the descriptors are normalized to be in the range $[0, 1]$.

The number of detected interest regions depends on the image. Therefore, an intermediary step to standardize the features within the images is required. This task is performed in the three next blocks. In the "Clustering" block, K-means algorithm is used to find clusters between patches and compute a set of centroids (*visual words*) called visual dictionary. This dictionary is constructed using all the patches of the training images. The size of the dictionary influences the performance of BoF. Therefore, different sizes are tested $K \in \{100, 200, 300\}$.

In the "Features quantization" block each image is separately analyzed and its corresponding patches are compared with the dictionary, thus it is possible to associate a *visual word* to each of the patches. The occurrence of each *visual word* in a lesion can be counted and a histogram of *visual words* frequency is computed for that lesion. This histogram will be a descriptor of the lesion, i.e. a feature vector, and will be fed to the classifier in the "Classification" block. Different classification algorithms can be used in the final block. In this paper, the classification rule is obtained using the k-Nearest Neighbor (kNN) algorithm. The parameters for this classifier (number of neighbors k and comparative distance) are optimized. Three different distances are used {Euclidean, Kolmogorov, Kullback-Leibler} and k is searched in the interval $\{5, 7, ..., 25\}$.

3 Sampling Methods

Skin lesions have different forms and aspects. Therefore, finding informative points and support regions that correctly describe a lesion is not a trivial task. A simple way of detecting the informative key points is assuming that each keypoint is a node of a regular grid placed on the lesion [9]. The interval between points is fixed and the patches extracted around each one of them have a size $\delta_d \times \delta_d$, $\delta_d \in \{20, 40, ..., 80\}$. In this work it is assumed that the interval between two consecutive nodes is δ_d to prevent patch overlapping. Fig.2(b) shows an example of the dense sampling method using δ_d=40.



a) b)δ_d=40

c)δ_s=6 d)δ_s=6 e)δ_s=6

Fig. 2. Examples of sampling strategies: a) Original Image; b) Dense Sampling; c), d) and e) Sparse Sampling: c) DoG Detector, d) Harris Laplace Detector and e) Hessian Laplace Detector

Dense sampling saves most of the information regarding the appearance of a lesion, hence has a great discriminative power. However, if a lesion is more or less homogeneous regarding color and texture, computational time will be lost on processing similar patches, which might be undesirable and redundant [9]. An alternative to this method is sparse sampling that extracts only the most informative regions. This strategy consist in using a salient point detector that searches for specific texture patterns like edges or blobs at different scales. Using different scales is a key characteristic of sparse sampling, since it makes the detected patches scale-invariant [9]. This is particularly relevant in dermoscopy since the same dermoscopic feature might appear in two different lesions at different scales. Although it is also possible to use different scales in the dense sampling method, this is both memory and time consuming.

A sparse sample method works as follows. First, a scale-space representation is constructed by convolving the image $I(x, y)$ with the Gaussian kernel $G(x, y, \sigma_D)$ at different scales σ_D. The second step is to detect the salient points and their characteristic scales. Several detectors that focus on different texture properties can be used to this aim. These detectors use the information of the scale-space and alter its representation using a specific function. Due to the different properties of the state-of-the-art descriptors, this work compares three of the most popular.

Difference of Gaussian (DoG): This detector was proposed by Lowe [12]. It computes the absolute diference between consecutive levels of the scale-space representation using the following function

$$D(x,y,\sigma_D) = |I(x,y) * G(x,y,\sigma_{D+1}) - I(x,y) * G(x,y,\sigma_D)| \qquad , \tag{1}$$

where σ_{D+1} and σ_D are two consecutive scales. The focus of the DoG detector are blob like structures. Therefore, this detector can be used to identify saliency points associated with circular dermoscopic features like dots and globules. A point is classified as saliency point if it is larger than its eight neighbor pixels in the corresponding $D(x,y,\sigma_D)$ and than the nine neighbors in the scales above and below. Fig.2(c) shows the keypoints detected using DoF and their characteristic scales.

Harris-Laplace: The Harris Laplace detector [13] identifies corner-like regions, thus it can be used to detected keypoints related with the presence of pigment network or streaks both of them dark linear structures. As in the previous detector, the search for the keypoints is performed by determining the 3D extrema. However, in this method a point and a scale are only selected if they verify the extrema condition for both Harris function and Laplacian operator. The Harris measure, responsible for detecting points in the scale-space, is the following

$$C(x,y,\sigma_s) = det(M(x,y,\sigma_D)) - \alpha trace^2(M(x,y,\sigma_D)) \qquad , \tag{2}$$

where α is a constant set to 0.06, as suggested in [13], and

$$M(x,y,\sigma_D) = \sigma_D^2 G(x,y,\sigma_I) * \begin{bmatrix} L_x^2(x,y,\sigma_D) & L_xL_y(x,y,\sigma_D) \\ L_xL_y(x,y,\sigma_D) & L_y^2(x,y,\sigma_D) \end{bmatrix} . \tag{3}$$

L_i denotes the first order i derivative of $I(x,y) * G(x,y,\sigma_D)$ and σ_I is the integration scale used to average the derivatives in the neighborhood of the pixel (x,y). The Laplacian operator,

$$Lap(x,y,\sigma_D) = |\sigma_D^2(L_{xx}(x,y,\sigma_D) + L_{yy}(x,y,\sigma_D))| \qquad , \tag{4}$$

is used for selecting the characteristic scale for each point. An example of the detected keypoints and their best scales is shown on Fig.2(d).

Hessian-Laplace: This detector follows an idea similar to the previous one. However, instead of using the Harris measure, the detection is performed using the scale normalized determinant of the Hessian matrix

$$H(x,y,\sigma_D) = \begin{bmatrix} L_{xx}(x,y,\sigma_D) & L_{xy}(x,y,\sigma_D) \\ L_{xy}(x,y,\sigma_D) & L_{yy}(x,y,\sigma_D) \end{bmatrix} \qquad , \tag{5}$$

that responds to blobs and ridges and penalizes elongated structures [14]. Therefore, this detector can be used to detect dots and other circular-shape dermoscopic structures. As before, the Laplacian operator is used for scale selection [14]. Fig.2(e) shows an example of the Hessian-Laplace detector.

The number of detected keypoints can influence the performance of BoF. Although a large number of keypoints allows a good description of a lesion, the computational time required for processing them will be great. Moreover, having a large amount of keypoints per image might not lead to the the creation of a generalized dictionary and a discriminative classification rule. Therefore, it is necessary to control the number of keypoints per image. This can be done by discarding the keypoints $(\bar{x}, \bar{y})$ that do not fulfill the following condition

$$S(\bar{x}, \bar{y}, \sigma_D) > Th.S_{max}(x, y, \sigma_D)) \tag{6}$$

where S is one of the detector functions and Th is a threshold found experimentally to be searched in the interval $Th \in \{0.01, 0.05, 0.1, 0.2\}$.

After detecting the keypoints it is necessary to extract their support regions. It is desirable that these patches are related with the characteristic scales of the keypoints. Therefore, for each keypoint it is extracted a patch of size $\sigma_D\delta_s \times \sigma_D\delta_s$, where $\delta_s \in \{4, 6, 8, 10\}$ in the case of color moments descriptors and $\delta_s \in \{8, 10\}$ for color histograms. The support regions for each type of detector are exemplified in Fig.2 using δ_s=6.

4 Results and Discussion

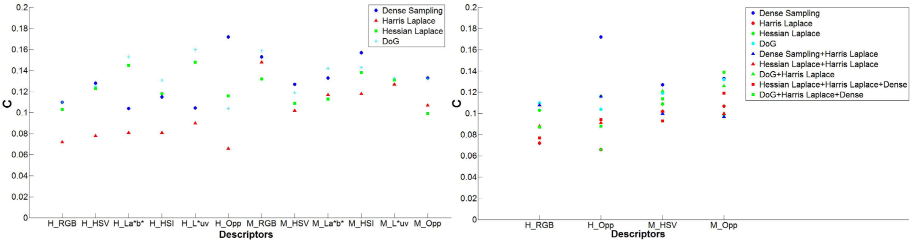
The detectors were tested on a dataset of 176 dermoscopy images of melanocytic lesions of which 25 were melanomas. These images were acquired during routine clinical exams in Hospital Pedro Hispano, Matosinhos, using a digital acquisition system with a magnification of 20×. The images were manually segmented and pre-processed using the algorithms described in [4]. An experienced dermatologist corrected the segmentations and classified the lesions as melanoma or not for a ground truth label.

The metrics used to determine the performance of the descriptors are Sensitivity (SE) and the Specificity (SP). A cost function ($\mathcal{C}$) that takes into account the trade-off between SE and SP is used to select the best results for each detector.

$$\mathcal{C} = \frac{c_{10}(1 - SE) + c_{01}(1 - SP)}{c_{10} + c_{01}} \quad , \tag{7}$$

c_{10} and c_{01} are the costs of an incorrectly classified melanoma and non-melanoma, respectively. An incorrectly classified melanoma is a worse error. Therefore, it was defined that c_{10} should be more penalizing than c_{01}. It was found that setting $c_{10} = 1.5c_{01}$ and $c_{01} = 1$ led to a good trade-off between SE and SP, i.e., achieving a high SE without significantly reducing the value of SP .

A 10-fold stratified cross validation method was used to compute the evaluation metrics. The set of images was divided in ten subsets, each with approximately the same size and number of positive examples. From these ten folds, nine were used for training and the remaining one for testing. This was repeated ten times each time with a different combination of sets for training and testing. The final results are the average of the ten training-testing processes. To



a) b)

Fig. 3. Classification results for a) Individual detectors and b) Detector fusion. The Y axis is the cost value C computed using (7).

deal with the problem of class unbalance, the number of positive local feature vectors was repeated until there was the same number of positive and negative examples. To avoid having several equal feature vectors in the training set, to each of the repetitions was added Gaussian Noise $N(0, \sigma_n^2)$, $\sigma_n = 0.0001$. This repetition was necessary since the smaller number of melanoma patches led to the construction of a poorly discriminative *visual dictionary.*

Fig.3(a) shows the results for the tested sampling strategies and detectors using different color descriptors. These results show that globally sparse sampling achieves better performances than dense sampling. This might be explained by the fact that dense sampling provides both relevant and irrelevant information for the dictionary construction and posterior training of the classifier, which lead to an incorrect classification. On the other hand, sparse sampling provides only regions of interest. The sparse detectors tested in this work use texture information to find the keypoints. Therefore, achieving better results with these detectors suggest that texture information, i.e. dermoscopic structures, plays a role on the characterization of melanomas. Regarding sparse detectors, Harris Laplace that looks for long structures performs better than DoG and Hessian Laplace. This suggests that linear dermoscopic structures might be more informative alone than circular structures. Table 1 shows the best results achieved with each of the sampling methods.

Table 1. Best classification results and descriptors for each sampling strategy

Sampling	Keypoint Detector	Descriptor	SE	SP	$\mathcal{C}$
Dense Sampling	Regular Grid	$h_{La^*b^*}$	93%	85%	0.104
Sparse Sampling	Harris Laplace	h_{Opp}	98%	86%	0.066
	Hessian Laplace	M_{Opp}	96%	81%	0.099
	Difference of Gaussian	h_{Opp}	98%	78%	0.104

To determine if the results could be further improved the best overall pairs of detectors/descriptors (h_{RGB}, h_{Opp}, M_{HSV} and M_{Opp}) are combined. This combination is done by late fusion [15], i.e., the final classification is a fusion of the output of different classifiers. In this work, the sum-rule is used to combine the outputs of the different classifiers using their respective posterior-probabilities [15]. The posterior probabilities for kNN are computed as follows

$$P(w|x) = \frac{k_w}{k} \qquad , \tag{8}$$

where w represents the class that can be either 0 or 1, x is a pattern to be classified and k_w is the number of patterns amongst the total number of neighbors k that belongs to class w. Fig3(b) shows the results for the combinations.These results suggest that the final classification improves if more than one type of keypoints detector is used. The best results were achieved with the fusion of the blobs detector DoG and the lines detector Harris-Laplace using h_{Opp}: SE=98%, SP=86%. In future works this combination of detectors and descriptor should be considered, since it allows the usage of both texture and color information in the classification of dermoscopy images.

5 Conclusions

BoF is a pattern recognition tool that can be used with success in the classification of dermoscopy images and that can provide relevant medical information. Understanding the role of the different keypoints detection methods was the aim of this paper. The achieved results suggest that a sparse sampling strategy performs better than a dense sampling strategy, achieving SE=98% and SP=86% with both the Harris Laplace detector and the combination of Harris Laplace and DoG detectors.

Future work should rely on studying the other steps of BoF and trying to retrieve relevant medical information from the results and dictionary.

Acknowledgments. This work was supported in the scope of 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. 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)
3. 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)
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. Celebi, M., et al.: Automatic detection of blue-white veil and related structures in dermoscopy images. CMIG 32(8), 670–677 (2008)
6. Serrano, C., et al.: Pattern analysis of dermoscopic images based on markov random fields. PR 42, 1052–1057 (2009)

7. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. 9th IEEE ICCV, pp. 1470–1477 (2003)
8. Situ, N., et al.: Evaluating sampling strategies of dermoscopic interest points. In: Proc. 8th ISBI, pp. 109–112 (2011)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part IV. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
10. van de Sande, K., et al.: Evaluating color descriptors for object and scene recognition. IEEE TPAMI 32, 1582–1593 (2010)
11. Yu, H.: Li, et al.: Color texture moment for content-based image retrieval. In: Proc. IEEE ICIP, vol. 3, pp. 929–932 (2002)
12. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. Jour. Comp. Vis. 60(2), 91–110 (2004)
13. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. Jour. Comp. Vis. 60(1), 63–86 (2004)
14. Mikolajczyk, K., et al.: A comparison of affine regions. Int. Jour. Comp. Vis. 65(1/2), 43–72 (2005)
15. Kittler, J., et al.: On combining classifiers. IEEE TPAMI 20, 226–239 (1998)