Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2011 December 31; 2011: 109–112. doi:10.1109/ISBI.2011.5872366.

EVALUATING SAMPLING STRATEGIES OF DERMOSCOPIC INTEREST POINTS

Ning Situ¹, Tarun Wadhawan¹, Rui Hu², Keith Lancaster², Xiaojing Yuan³ [IEEE Member], and George Zouridakis^{1,2,3} [IEEE Senior Member]

¹Department of Computer Science, University of Houston, Houston, TX 77204, USA

²Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA

³Department of Engineering Technology, University of Houston, Houston, TX 77204, USA

Abstract

Among the most critical components of a computerized system for automated melanoma detection is image sampling and pooling of the extracted features. In this paper, we propose a new method for sampling and pooling based on a combination of spatial pooling and graph theory features. The performance of the new method is evaluated using a dataset of more than 1,500 images representing pigmented skin lesions of known pathology. In our comparisons, we include several methods ranging from simple and multi-scale sampling on a regular grid to more sophisticated approaches, such as blob and curvilinear structure detectors. Our results show that, despite its simplicity, simple sampling on a regular grid provides highly competitive performance, compared to the more sophisticated approaches, while multi-scale sampling yields only trivial improvements. However, the proposed method provides significant performance improvement in terms of sensitivity and area under the receiver operating characteristic curve (95% t-test), and the best performance in terms of specificity compared to all other methods explored.

Keywords

Image classification; interest point detection; dermoscopic image; pigmented skin lesion

1. INTRODUCTION

The American Cancer Society predicts that there will be approximately 68,130 new cases of melanoma and 8,700 deaths because of melanoma in US in 2010 [1]. Thus, detection of an early melanoma is of paramount importance for successful skin cancer screening. The use of Dermoscopy, an imaging technique to visualize structures inside pigmented skin lesions beyond the naked eye, and computerized systems for automated classification of dermoscopic images [2] can drastically improve the diagnostic accuracy of early melanoma.

Image classification using interest points has shown success in previous studies [3]. Of particular importance is the concept of dermoscopic interest points (DIPs, [4]) which represent crucial dermoscopic features for accurate diagnosis. Zhou *et al.* [4] proposed the use of a blob detector and a curvilinear structure detector for DIPs and compared the results obtained with regions of interest drawn by domain experts. He also tested the invariance property of different descriptors and samplers with respect to scaling, orientation, and illumination. However, the discriminative property of DIP was not evaluated.

We have recently developed an automated procedure for lesion classification [5] based on the well known bag-of-features (BoF) framework [3]. In the present study we use the same BoF framework to implement lesion classification using DIPs. Two of the most important steps for correct classification are image sampling and feature pooling. At present, most methods use average pooling [3, 5] inside the whole image region, which unfortunately, does not consider any spatial information. Thus, we propose a new method that combines a spatial pooling scheme with graph theory features.

The new method is evaluated using a set of more than 1,500 images of known classification and its performance, in terms of classification of a lesion as malignant or benign, is compared against several other methods for sampling and pooling. In this study the benign category includes melanocytic and nonmelanocytic benign lesions as well as Clark or dysplastic nevi.

2. METHODS

Our lesion classification procedure [5] consists of five main steps: image sampling, feature extraction, coding, pooling, and final classification [6]. For a given image, we first identify DIPs inside the lesion and then extract a patch around each DIP. On each patch, we then compute several low level texture and color features using Haar wavelets and color moments, which are important for melanoma detection. In the coding stage, a patch is assigned to a codeword from a pre-learned codebook using hard or soft assignment. In our case, each patch is assigned to its nearest neighbor in the codebook with hard assignment. We then pool the assignments of all patches extracted from a lesion into one feature vector. The last step is to classify the lesion based on the feature vector obtained from pooling. The following analysis focuses on evaluating only the image sampling and feature pooling components of the whole system and their effects on the final classification of a lesion.

2.1. Image Sampling Strategies

The sampling operator selects N pixels inside a lesion and then it centers a $p \times p$ pixel patch at each pixel location. For DIP detection, we investigate four sampling strategies: The first two are specifically designed for blobs and curvilinear components, respectively, which are the typical structures seen inside a lesion [4]. The other two, however, are not targeting any particular lesion structure; yet, they result in excellent image classification performance.

Detector for blobs and corners—Blobs, dots, and globular structures are frequently observed in a lesion. We use the scale-invariant feature transform (SIFT) [7] to detect these structures, a procedure also used in [4]. Fig. 1 gives an example of a SIFT detector.

Detector for curvilinear structures—the SIFT operator is not stable for ridge detection [7] and it may fail to localize curvilinear structures in the lesion, as it was noted also by Zhou *et al* [4]. Instead, for curvilinear structures, we apply a Frangi filter (Frangi) [9] at three scales $\sigma=1, 2$, and 3 [9]. Points with higher filter responses have higher probabilities of being curvilinear structures. A Frangi filter is similar to the Steger filter [10] used in [4]. An example of this procedure is shown in Fig. 2.

Grid sampling—Sampling on a regular grid of size of g (Grid- g) placed on a lesion. When g is small, this is also called dense sampling.

Radial sampling—Sampling using polar coordinates on axes placed on the lesion with origin at the center of the lesion (Radial). The rationale behind this scheme is that a lesion generally follows a radially growing pattern [11].

2.2. Feature Pooling Schemes

We investigate two pooling schemes: the popular average pooling and a spatial pooling.

Average pooling uses averaging of the class assignments across all patches. This is equivalent to building a normalized histogram, whereby each bin corresponds to a codeword in a codebook and the bin's value is proportional to the number of patches assigned to that codeword.

Spatial pooling detects homogeneous regions inside a lesion and then uses average pooling in each homogeneous region. A lesion is segmented into 3 to 8 regions using the normalized cut method (*e.g.*, Fig. 3). Tiny regions are grouped with nearby larger ones. Thus, after spatial pooling, a single vector (histogram) is produced for each segmented region.

In the proposed method, a whole lesion is represented as a fully connected weighted graph, whose nodes correspond to homogeneous regions. The weight of an edge is the Euclidean distance between the vectors of the two connected nodes (regions). We then represent a lesion using six features implemented in the graph measure toolbox [12], namely clustering coefficient, maximized modularity, characteristic path length, eccentricity for each vertex, radius, and diameter of the graph (graph eccentricity, radius, and diameter are not the same lesion measures defined in [2]). Tree and graph schemes have been proposed before (*e.g.*, [13, 14]), however, not for malignant classification. The proposed weighted graph model extends our recent work in which we employed a non-weighted graph lesion representation for melanoma detection [15].

2.3. Codebook and Classifier Implementation Details

We build codebooks using K-mean clustering on a set of patches obtained by randomly sampling 1,000 patches from every lesion so that every lesion contributes equally to the codebook construction. Thus, our evaluation uses transductive inference [16] *i.e.*, in our classifier learning method we employ labeled training data and unlabeled testing data while for testing we predict labels for the latter. The number of cluster is 200 for wavelet features and 100 for color features. The overall performance is not sensitive to these choices. We build separate codebooks for wavelet and color, and different codebooks for the three patch sizes: 16, 24, and 32. By default, we use average pooling, if not specified otherwise. Our classifier uses support vector machines (SVMs) with a χ^2 kernel, which is the state-of-the-art setting for BoF model. For graph theory features, we use a Gaussian kernel which is the common choice for SVMs. We choose the threshold for the classifier's output by maximizing the average of sensitivity and specificity on the labeled training data. For classifier combination, we use simple ensemble averaging (weighted combination [17] yielded very similar results on our dataset).

2.4. Data and Procedure

We use a dataset with 1,505 epiluminescence microscopy (ELM) images, in which 1,098 lesions are benign and 407 lesions are melanoma. The images are mainly collected from [8, 18]. The image size ranges from 712×454 to 1,024×768 pixels, and lesion size ranges from 7,662 to 804,527 pixels. We use manual segmentation of all lesions to ensure that evaluation of the various sampling strategies is not affected by possible differences in automated identification of the lesion boundary.

We perform ten times ten-fold stratified cross-validation. We use sensitivity, specificity, balanced accuracy (BAC, *i.e.*, average of sensitivity and specificity), and area under the receiver operating characteristic curve (AUC) as performance criteria. For sensitivity and specificity, we report the mean and 95% confidence interval (CI) estimated from a binomial

distribution, and their average for BAC. For AUC we show both the mean value and standard deviation (std) of the values obtained.

3. RESULTS

3.1. The Effect of Number of Patches Sampled

Choosing the same number of patches for all lesions is not reasonable, since lesions differ in size. Instead, we choose a number of patches proportional to that lesion's area. We use simple grid sampling and grid size chosen from the set $\{1, 5, 10, 20, 40, 100\}$. Using a grid size g is equivalent to sampling approximately $(100/g^2)\%$ points from a lesion. We use a square patch of 24 pixels in size. Fig. 4 shows that this percentage value affects significantly both performance measures. BAC starts to converge when the number of points approaches 4% of the lesion's area, while AUC converges earlier at about 1%. Thus, we need to sample only 4% of points (*i.e.*, Grid-5) from a lesion without decreasing performance significantly (for both BAC and AUC).

3.2. The Effect of Sampling Strategy

We now consider four lesion sampling methods, namely Grid-5, Radial, SIFT, and Frangi, and adjust the parameters and thresholds of the latter three methods to retain 4% of all possible samples. In addition, we combine the classifiers from Radial, SIFT, and Frangi with Grid-1 (denoted as Com) to test whether classification accuracy improves when combining classifier training with interest points located at dermoscopic structures instead of simply using all possible points alone (*i.e.*, Grid-1). Fig. 5 shows that regular grid sampling Grid-5 provides results comparable to Radial, SIFT, and Frangi. A comparison between Com and Grid-1 reveals only a marginal improvement in BAC, but no improvement in AUC, when incorporating the more complicated interest point detectors instead of using simple dense sampling alone.

3.3. The Effect of Sampling at Multiple Scales

For each sampling strategy, we extract square patches of size 16, 24, and 32 pixels, and then we group the classifiers obtained from these three scales. For the multi-scale model of Com, we ensemble 12 classifiers from four sampling methods and three scales. Fig. 6 shows that multi-scale sampling can improve the performance of some methods compared to sampling at a single scale with patches of size 24. However, none of the multi-scale models in Fig. 6 is significantly better than Grid-1 using a single scale sampling.

3.4. The Effect of Spatial Pooling

We use spatial pooling for patches of size 16×16 centered on every pixel, since we empirically observed that a patch of size 16 performs better than size 24 or size 32 for graph theory features. We ensemble the classifiers built from spatial pooling and Grid-1, and denote the combined model as DenSpa. We compare DenSpa with Grid-1, Com, and the multi-scale models of Grid-1 and Com, denoted as GridMul and ComMul, respectively, in Table 1. We see that DenSpa performs the best among the five schemes in all measures. The mean sensitivity of the other four methods without spatial pooling is below the 95% CI of DenSpa. The improvement for specificity is not so significant, but the AUC of DenSpa is significantly different from the other four methods as revealed by an unpaired t-test at 95% confidence level.

4. DISCUSSION

Our results show that dense sampling on a single scale with average pooling performs very well when compared against three other sampling schemes, namely SIFT, Frangi, and Radial, or multi-scale sampling. Furthermore, our results demonstrate that spatial pooling can improve overall accuracy of melanoma detection. One possible explanation is that malignant lesions usually exhibit high heterogeneity in texture and/or color, and the use of graph theory features on a weighted graph built from spatial pooling can capture this heterogeneity.

Grid-5 uses only about 4% of the patches used by Grid-1, which extracts all possible patches of size 24 from a lesion; yet, the performance of the former is similar to that of the latter. A sparser representation of a lesion has the advantage of reduced computational time. For instance, Grid-5 is theoretically 25 times faster than Grid-1. This can be significant when computational resources are limited, such as, for example, when screening of melanoma using a handheld device.

5. CONCLUSIONS

Evaluation on a dataset with more than 1,500 lesions shows that sampling on a regular grid can serve as a highly competitive baseline for sampling strategies of dermoscopic interest points for the purpose of classification between melanoma and benign lesions. We find that combining spatial pooling and graph theory features provides a significant performance improvement for sensitivity and area under the receiver operating characteristic curve (95% t-test).

Acknowledgments

This work was supported in part by NIH grant no. 1R21AR057921, and by grants from UH-GEAR and the Texas Learning and Computation Center at the University of Houston.

6. REFERENCES

- [1]. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA: a cancer journal for clinicians. 2010; vol. 60(no. 5):277. [PubMed: 20610543]
- [2]. Ganster H, Pinz A, Rohrer R, Wildling E, Binder M, Kittler H. Automated melanoma recognition. IEEE Transactions on Medical Imaging. 2001; vol. 20(no. 3):233–239. [PubMed: 11341712]
- [3]. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. ECCV; Workshop on Statistical Learning in Computer Vision; 2004. p. 22
- [4]. Zhou, H.; Chen, M.; Rehg, JM. Dermoscopic interest point detector and descriptor. Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on; IEEE. 2009. p. 1318-1321.
- [5]. Situ, N.; Yuan, X.; Chen, J.; Zouridakis, G. Malignant Melanoma Detection by Bag-of-Features Classification. Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2008. p. 3110
- [6]. Yang, J.; Yu, K.; Gong, Y.; Huang, T. Linear spatial pyramid matching using sparse coding for image classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE. 2009.
- [7]. Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. 2004; vol. 60(no. 2):91–110.
- [8]. Argenziano G, Soyer HP, De Giorgi V, Piccolo D, Carli P, Delfino M, Ferrari A, Hofmann-Wellenhof R, Massi D, Mazzocchetti G, et al. Dermoscopy: a tutorial. 2000; vol. 12
- [9]. Frangi AF, Niessen WJ, Vincken KL, Viergever MA. Multiscale vessel enhancement filtering. Medical Image Computing and Computer-Assisted Intervention-MICCAI 98. 1998:130.

- [10]. Steger C. An unbiased detector of curvilinear structures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2002; vol. 20(no. 2):113–125.
- [11]. Zhou, H.; Chen, M.; Gass, R.; Ferris, L.; Drogowski, L.; Rehg, JM. Spatially constrained segmentation of dermoscopy images. *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*; IEEE. 2008. p. 800-803.
- [12]. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*. 2009
- [13]. Cucchiara R, Grana C. Using the Topological Tree for skin lesion structure description. *Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies: Kes 2002*. 2002:166.
- [14]. Sadeghi M, Razmara M, Ester M, Lee TK, Atkins MS. Graph-based pigment network detection in skin images. *Proceedings of SPIE*. 2010; vol. 7623:762312.
- [15]. Situ, N.; Wadhawan, T.; Yuan, X.; Zouridakis, G. Modeling Spatial Relation in Skin Lesion Images by the Graph Walk Kernel. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.; 2010.
- [16]. Vapnik, V. *Statistical learning theory*. Vol. vol. 2. Wiley; New York: 1998.
- [17]. Lanckriet GRG, Cristianini N, Bartlett P, El Ghaoui L, Jordan MI. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*. 2004; vol. 5:27–72.
- [18]. Menzies, SW.; Crotty, KA.; Ingwar, C.; McCarthy, WH. *An atlas of surface microscopy of pigmented skin lesions: dermoscopy*. McGraw-Hill Book Company; Australia: 2002.

Fig. 1.
Example of SIFT detector: original image from [8] (left) and interest points (right) detected by SIFT using a threshold adjusted to retain 0.25% of points in the lesion.

Fig. 2.

Example of Frangi filtering: original (left, from [8]) and filtered (right) image. Higher responses (brighter spots) are seen at curvilinear structures in the periphery of the lesion.

Fig. 3.
Segmentation inside the lesion in Fig. 1.

Fig. 4. Effect of the total points sampled on classification accuracy. PSL: pigmented skin lesion.

Fig. 5.
Comparison of various sampling schemes for 24x24 patches.

Fig. 6.
Single scale compared to multi-scale sampling.

Table 1

Classification performance of five schemes.

Methods	Sensitivity (95%CI)	Specificity (95%CI)	AUC (std)
Grid-1	82.49 [79.16, 85.40]	83.32 [81.39, 85.07]	90.93±2.67
GridMul	82.31 [78.92, 85.18]	84.15 [82.28, 85.88]	91.16±2.55
Com	81.96 [78.67, 84.97]	84.68 [82.82, 86.37]	90.87±2.60
ComMul	82.40 [79.16, 85.40]	84.19 [82.37, 85.97]	90.99±2.53
DenSpa	86.17 [83.11, 88.80]	84.68 [82.82, 86.37]	92.71 ±2.25