Malignant Melanoma Detection by Bag-of-Features Classification

Ning Situ, Xiaojing Yuan, *IEEE Member*, Ji Chen, *IEEE Member*, and George Zouridakis, *IEEE Senior Member*

***Abstract*— In this paper, we apply a Bag-of-Features approach to malignant melanoma detection based on epiluminescence microscopy imaging. Each skin lesion is represented by a histogram of codewords or clusters identified from a training data set. Classification results using Naive Bayes classification and Support Vector Machines are reported. The best performance obtained is 82.21% on a dataset of 100 skin lesion images. Furthermore, since in melanoma screening false negative errors have a much higher impact and associated cost than false positive ones, we use the Neyman-Pearson score in our model selection scheme.**

I. INTRODUCTION

Local patterns are important features for early melanoma detection. Many criteria like the ABCD rule, 7-point checklist, and Menzies' method used by dermatologists are based on the presence of certain texture patterns. Most existing studies focus on detecting a specific texture pattern, such as a dark area by Pellacania et al. [6], asymmetric blotches by Stoecker et al. [9], and irregular streaks and atypical pigmented network by Betta et al. [2]. In this paper, we attempt to build classifiers for melanoma detection based on the distribution of local patterns. Bag-of-Features based image classification is widely used in computer vision [3], [4]. It is analogous to Bag-of-Words for document modeling, and models an image as a histogram of "visual words" that serves as an input feature vector for the classification algorithm. "Visual words" in a codebook are built from quantization of descriptors of local image patches which are sampled from a training set. Quantization can be performed by standard clustering algorithms, such as k-mean and EM, and the centroid of each cluster is a "visual word" in the codebook. A new image is represented by image patches, and each patch is assigned to the cluster of its nearest neighbor in the codebook.

A codebook is used to quantize the continuous signal into a discrete one [3]. We can also cluster the local image patches into groups to discover similar skin lesion patterns among the training data. This is analogous to discovering common "topics" shared among several documents by grouping "words" into clusters. One problem here is that we do not know the number of clusters or the number of different patterns present in the training images. Dirichlet process is a recently used method to choose the number of cluster. Orbanz et al. [5] accomplished image segmentation based on a Dirichlet process that incorporates Markov Random Field. We applied their algorithm to discover the shared clusters among skin lesion images. As before, the histogram of the "topics" can also be the "signature" of a skin lesion. Controlling the sensitivity and specificity of the classifier is crucial. The "2C-SVM" algorithm [10] is applied to generate a classifier by placing different weights on positive and negative samples.

Since in cancer screening the cost of a false negative error is much higher than a false positive, we require that the false negative rate be smaller than some threshold, while minimizing the false positive rate. To achieve this objective, a Neyman-Pearson score (NPS) has been proposed by Scott [8] for model selection and NPS can be used for any learning algorithm. Scott [8] proved that with sufficient large training samples, the model selected by the NPS criterion can ensure that the false negative rate on test data will remain below a certain threshold. ROC curve and AUC criteria are also widely used for model selection, but as pointed out by Scott [8], ROC measures compare the performance of a family of classifiers. Instead, our ultimate goal is to find a single classifier for our automated skin-cancer detection system, and Neyman-Pearson score seems the proper choice.

II. METHODS

A. Bag-of-Features

Each skin lesion is represented by a Bag-of-Features defined on several patches sampled on the image. To describe each patch, we used wavelets and "Gabor-like" [7] filters in our experiment, but several other texture features can also be used. A 3-level wavelet decomposition is applied to 16×16 image patches and the energies of the 10 subbands are used as patch descriptors. The advantage of the "Gabor-like" [7] filters developed by Schmid is that they are invariant to rotation. The energies of 13 channels are used as descriptors for one image patch and the features are normalized by the mean and variance estimated from the training data [7]. K-means clustering is used to quantize the features.

B. Mixture Dirichlet Process and Markov Random Field

The Mixture Dirichlet process (MDP) has recently become a popular method to choose the number of clusters. Orbanz and Buhmann [5] incorporated Markov Random Field (MRF) into MDP and developed the combined MDP/MRF segmentation method. Their idea can be explained by the "Chinese Restaurant Process" (CRP): suppose that there are n data

This work was supported in part by NSF grant no. 521527, and grants from UH GEAR, ISSO, the Texas Learning and Computation Center, and the Texas Heart Institute.

X. Yuan (phone: 713-743-1129; fax: 713-743-4032; xyuan@uh.edu), G. Zouridakis, and J. Chen are with the Departments of Engineering Technology, Computer Science, and Electrical and Computer Engineering, University of Houston

N. Situ is a Ph.D. candidate in Computer Science at the University of Houston.

points (image patches in our problem), and $(x_1, x_2, \ldots, x_n)$ denote the cluster of the corresponding patches. Let n_k denote the number of patches in cluster k and c the number of clusters, and let x^{-i} and n_k^{-i} indicate that we exclude the ith patch. By assuming exchangeability of every two patches, a nonparametric prior is defined as follows:

$$P(x_i = k|x^{-i}) \propto n_k^{-i}\text{exp}(-H(k|x^{-i})) \quad (1)$$

$$P(x_i \neq k, k = 1, 2, \ldots, c|x^{-i}) \propto \alpha \quad (2)$$

The term $H(k|x^{-i})$ will have a smaller value if the neighbor of patch i belongs to cluster k (8-connected neighborhoods are used in our experiment). The prior thus defined has the following meaning: assuming that we know the assignment of all patches except i, the probability of patch i belonging to cluster k is proportional to the number of patches in cluster k times the effect of the term H. The effect of the term H above is obvious: neighboring patched are encouraged to be grouped into the same cluster. The probability of patch i belonging to a new cluster is proportional to α. If patch i belongs to a new cluster, it draws a new parameter from a base measure G_0 for that new cluster. A detailed math derivation can be found elsewhere [5].

To make an inference, a Gibbs sampling algorithm has been developed [5]. The parameters of each cluster can also be obtained by running the Gibbs sampling algorithm on training images. For a new image, the same sampling scheme from [5] can be used, except that the parameters are fixed to those learned from a training set instead of updating them at each step. Following [5], the features extracted for each patches are quantized local histograms of intensity (8 bins quantization is used in our experiments).

C. Classifiers

Similar to previous work [3], two types of classifiers are employed in this study: Naive Bayes classifier and Support Vector Machines (SVM). In order to control the sensitivity and specificity of SVM, the so-called "2C" formulation of SVM described below is used in our experiment to generate ROC curves and control false negative error rate.

1) 2C Support Vector Machines (SVM): Assume that we have m benign samples $(x_1, y_1), \ldots, (x_m, y_m)$ and n malignant samples $(x_{1+m}, y_{1+m}), \ldots, (x_{m+n}, y_{m+n})$, where x_i is the feature vector of lesion i, i.e. the histogram of lesion i, and y_i is the corresponding label. The "2C" SVM algorithm proposed by Veropoulos et al. [10] is formulated as follows:

$$\min_{w,b,\xi} \quad \frac{1}{2}|w|^2 + C_1(C_2\sum_{i=1}^{m}\xi_i + (1-C_2)\sum_{i=1+m}^{m+n}\xi_i) \quad (3)$$

$$\text{s.t.} \quad y_i(w^T \cdot \phi(x_i) + b) \geq 1 - \xi_i, i = 1, 2, \ldots, m+n \quad (4)$$

$$\xi_i \geq 0, i = 1, 2, \ldots, m+n \quad (5)$$

where C_1 represents the trade-off between the regularization term and the empirical cost estimated from the training set, while C_2 is a parameter in $(0, 1)$ representing the weights of the two classes. ROC curves can be obtained by changing the value of C_2 within $(0, 1)$.

D. Neyman-Pearson Score for Model Selection

ROC analysis compares a family of classifiers rather than giving a particular classifier. In addition, for cancer detection, controlling the sensitivity is more important. In other words, we want to control the false negative rate (P_{FN}) within a certain range ($\leq \alpha$) while minimizing the false positive rate (P_{FP}). In practice, however, a strict criterion $P_{FN} \leq \alpha$ may not be feasible. It depends on the training data and the complexity of the model designed. Hence, in practical implementations, $P_{FN} \leq \alpha + \epsilon$ is used instead. Such a learning paradigm is called Neyman-Pearson learning [8], and Scott [8] proposed the Neyman-Pearson score (NPS) both for model and parameter selection,

$$NPS = \frac{1}{\alpha}\text{max}(\widehat{P_{FN}} - \alpha, 0) + \widehat{P_{FP}}$$

where $\widehat{P_{FN}}$ and $\widehat{P_{FP}}$ are empirical estimates from training data. The advantage of this criterion is that it can be used for any kind of learning algorithm.

III. EXPERIMENTS

A. Experiment Setup

Algorithms are tested on a data set consisting of 100 epiluminescence microscopy skin-lesion images, 70 images are benign and include nevocellular nevi and benign dysplastic nevi. The other 30 skin lesion images are malignant melanomas. Data are randomly split into 80% for training and 20% for testing. The ratio between benign and malignant images is maintain the same both in the training and testing sets, while an 8-fold cross validation is employed for model and parameter selection. Results are reported as the average of 100 repeated experiment. Skin lesions are segmented by active contour methods developed in our lab [11]. Segmentation results are validated by manual segmentation from three dermatologists [11]. A bounding box of the segmented lesion is extracted and scaled to 256×256 pixels for further processing, and 16×16 patches are sampled from a 16×16 regular grid placed on the 256×256 ROI. Patches whose area is more than 50% outside the skin lesion are discarded.

B. Codebook and Shared Cluster

Wavelet and "Gabor-like" filters are applied to each 16×16 image patch. A set of 10 features are obtained from the wavelet filter and 13 more are obtained from the "Gabor-like" filter. A widely used method for quantization is k-means clustering. A universal way to determine the codebook size has not been developed yet. It is observed that larger codebook sizes can lead to obtain higher accuracy [3], [4]. But, overfitting is also reported in [4] with a large codebook size. In our experiments, we tested the effect of different codebook sizes, namely $16, 32, 64, 128, 256$, and 512.

In accordance with the combined MDP/MRF method developed previously [5], we set $\lambda = 5$ (higher values result in a stronger smoothing effect), $\alpha = 10^{-4}$ (the recommended values are between 1 and 10^{-5}), while 400 iterations are used for the Gibbs sampler. The typical number of cluster we obtained in 100 repeated experiments ranged

from 38 to 41 by clustering more than 13000 patches. Fig. 1 shows an example from the 100 experiments performed. Two malignant lesions are shown in (a) and (b). The clustering algorithm automatically discovered shared clusters in the two lesions which are shown in (c) and (d). We can see that the shared cluster corresponds to the so called dark area [6] which is usually present in malignant lesion. Panels (e) and (f) are the histograms of clusters or the "signatures" of the two lesions, which will be the input of classification algorithm.

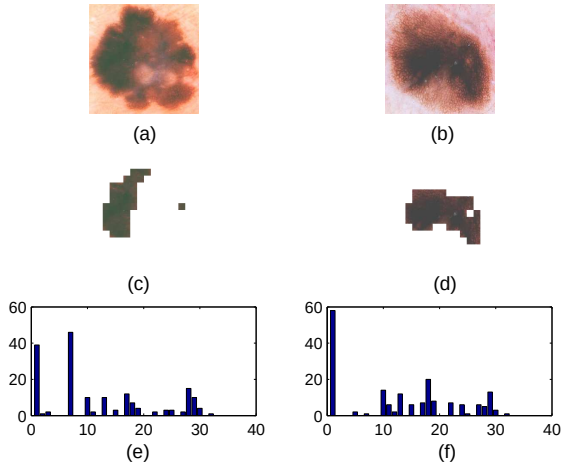
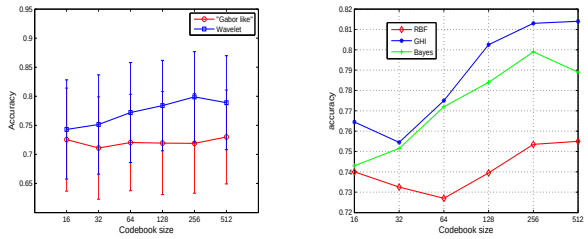


Fig. 1. (a) and (b) are two malignant skin lesion images. (c) and (d) are shared cluster(dark area blotches) discovered by the algorithm. (e) and (f) are histograms of clusters for the two skin lesions. The first cluster corresponds to (c) and (d).

C. Naive Bayes Results

We present classification results using features extracted from three methods: (1)wavelet and k-mean, (2)"Gabor-like" filters and k-mean, and (3) local histogram and MDP/MRF. Fig. 2(a) shows the classification accuracy results for codebooks built from wavelet features and "Gabor like" features with different codebook sizes. Fig. 2(a) clearly shows that



(a) (b)

Fig. 2. (a) Classification accuracy using "Gabor-like" feature (red) and wavelet feature (blue) with different codebook size; (b) Accuracy of SVM with GHI kernel(blue), RBF kernel(red), and Bayes (green) classifiers with different codebook size.

wavelet features outperform "Gabor like" features for different codebook sizes. Classification accuracy using features obtained from local histograms and combined MDP/MRF is only 76.85%, which does not outperform wavelet feature with a codebook size of higher than 64. Wavelet feature provides better results than "Gabor-like" feature and feature obtained from combined MDP/MRF on simple Bayes classifier. The rest of the experiments based on the more advanced SVM classifier are done using only wavelet features.

TABLE I
AUC FOR TWO TYPE OF KERNELS:RBF AND GHI.

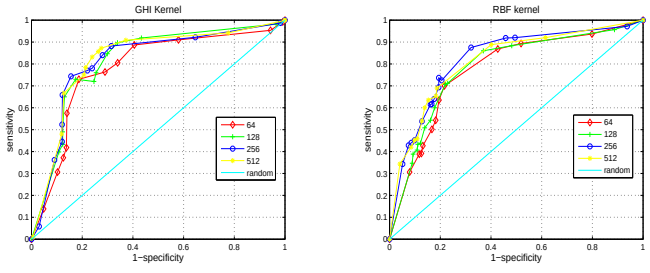
size	64	128	256	512
RBF	77.43%	78.36%	82.11%	81.29%
GHI	78.19%	81.64%	81.37%	**82.21%**

D. SVM Results

For standard "C-SVM", C is searched in the set $\{2^{-5}, 2^{-4}, \ldots, 2^{4}, 2^{5}\}$. Two types of kernel are used in our experiment: RBF kernel and Generalized Histogram Intersection kernel (GHI) [1]. In the former case, the width parameter is searched in the set $\{2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}\}$, whereas in the latter, Boughorbel et al.,[1] reported that the parameter β gives good result when near 0.25. Then, we search β in the set $\{0.2, 0.25, 0.3\}$. The accuracy results reported in Fig. 2(b) show that the GHI kernel is better than the RBF kernel and Naive Bayes classifier by using the histogram features extracted earlier.

E. SVM Model Selection by ROC Analysis

ROC curve and AUC (area under curve) are widely used measures to analyze the performance of classifiers. ROC curves for the two types of kernels RBF and GHI with different codebook sizes are shown in Fig. 3(a) and Fig. 3(b). The ROC curves are obtained by connecting sampled points, since it is time expensive to get all possible points. Then, the calculated values for the AUC corresponding to the curves thus obtained are only approximations, and they are reported in Table I. Again, we see that the GHI kernel outperforms the RBF kernel for various codebook sizes. A codebook size greater than or equal to 128 for the GHI kernel and 256 for RBF kernel can give good results. The best AUC's are obtained by a model with a codebook size of 512 that uses the GHI kernel.



(a) (b)

Fig. 3. (a) ROC curve of SVM with GHI kernel. (b) ROC curve of SVM with RBF kernel and codebook size 64 (red), 128 (green), 256 (blue), and 512 (yellow).

TABLE II

RESULTS OF MODEL SELECTION BY NPS FOR GHI KERNEL.

Parameters	sensitivity	specificity	NPS
$\alpha = 0.2$, size= 128	80.50%	74.07%	25.93%
$\alpha = 0.2$, size= 256	78.00%	77.75%	32.25%
$\alpha = 0.2$, size= 512	80.33%	78.50%	**21.50**%
$\alpha = 0.15$, size= 128	84.50%	71.93%	**31.40**%
$\alpha = 0.15$, size= 256	82.17%	74.79%	44.08%
$\alpha = 0.15$, size= 512	83.17%	76.29%	35.91%
$\alpha = 0.1$, size= 128	88.80%	69.29%	**42.71**%
$\alpha = 0.1$, size= 256	86.83%	71.57%	60.13%
$\alpha = 0.1$, size= 512	86.00%	72.21%	67.79%

F. SVM Model Selection by Neyman-Pearson Score

Instead of comparing a family of models, we are more interested in finding one single classifier. Model selection by the Neyman-Pearson score is suitable for this purpose [8]. C_2 for "2C SVM" is searched in the set $\{2^{-1}, 2^{-2}, \ldots, 2^{-15}\}$, while C_1 is the same as the C in standard SVM used before. Results of different controls ($\alpha = 0.2, 0.15, 0.1$) using the GHI kernel are reported in Table II. Sensitivity is successfully controlled to be above $1-\alpha$ when $\alpha = 0.2$ with a codebook size of 128 and 512. When α decreases to 0.15 and 0.1, the sensitivity also increases close to $1-\alpha$, but fails to meet the strict criteria, since for a smaller value of α, more training samples are required for proper model selection. With limited number of training samples, our results indicate that model selection by NPS is still effective to control the false negative error rate.

IV. DISCUSSION

The results presented above indicate that the GHI kernel is preferred to RBF kernel for the features we extracted, since the "signature" we used for each lesion is a histogram, and the GHI kernel is more suitable for classifying histogram data [1]. Choosing a proper kernel is important for the SVM to give good performance. The widely used RBF kernel performed even worse than the simple Bayes Classifier in Fig. 2(b).

Classification performance with features extracted from MDP/MRF and local histograms IS not improved compared to k-means with wavelet features and a large codebook size. However, the MDP/MRF method can provide meaningful clusters (Fig.1(c) and (d)) that are interpretable by humans.

In our experiments using Neyman-Pearson score for model selection, the α value was 0.2, 0.15, and 0.1. But in practical screening of malignant melanoma, a much smaller value of α should be used. A further decrease in the value of α causes the specificity to drop below 70%, which does not have much practical value. One reason for this is the limited size of the training data we used. To obtain a good result for smaller values of α a larger training sample would be required.

V. CONCLUSIONS AND FUTURE WORK

In this study, we present experimental results obtained by applying the Bag-of-Features approach to the problem of automatic detection of malignant melanoma. The "signature" of a skin lesion is obtained by building a codebook with texture features and k-means quantization. A method to discover shared clusters among lesions by the Dirichlet process has also been tested here. and the best classifier was obtained with an AUC of 82.21% from wavelet features and a codebook size of 512. Neyman-Pearson score is used to choose a single classifier and to control the false negative rate. Our experiment results demonstrate that model selection by Neyman-Pearson score is effective. To further improve classification performance, combining other features like color and border will be our future work. For the Dirichlet process, we only use local histogram features as [5]. Other texture features will be included into the MDP/MRF method and hierarchical models will also be explored. This requires a more efficient algorithm, as Gibbs sampling is too slow to converge. It is also interesting to see whether the clusters detected by computer algorithms match the patterns obtained when using the criteria established by dermatologists.

VI. ACKNOWLEDGMENT

We are grateful to Dr. William V. Stoecker, M.D., who provided the skin lesion dataset and the manual image segmentations from domain experts.

REFERENCES

[1] S. Boughorbel, J. Tarel and N. Boujemaa, "GENERALIZED HISTOGRAM INTERSECTION KERNEL FOR IMAGE RECOGNITION", In *Proceedings of the IEEE International Conference on Image Processing, 2005. (ICIP 2005)*, vol. 3, Sep. 2005.

[2] G. Betta, G. Leo, G. Fabbrocini, A. Paolillo and M.Scalvenzi, " Automated Application of the "7-point checklist" Diagnosis Method for Skin Lesions: Estimation of Chromatic and Shape Parameters. ", In *Proceedings of the IEEE Instrumentation and Measurement Technology Conference, 2005. (IMTC 2005.)*, vol. 3, 16-19 May 2005 pp. 1818 - 1822.

[3] S. Lazebnik and M. Raginsky, "Learning Nearest-Neighbor Quantizers from Labeled Data by Information Loss Minimization", In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, March 21-24, 2007.

[4] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification", *European Conference on Computer Vision(ECCV)*, p. IV: 490-503, 2006.

[5] P. Orbanz and J. M. Buhmann., "Nonparametric Bayesian Image Segmentation", In *International Journal of Computer Vision (IJCV)*, vol. 77, 2008, pp 25-45.

[6] G. Pellacania, C. Granab, R. Cucchiarab and S. Seidenaria, "Automated Extraction and Description of Dark Areas in Surface Microscopy Melanocytic Lesion Images", *Dermatology*, 208(1): pp 21-26, 2004.

[7] C. Schmid, "Constructing models for content-based image retrieval", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2001,pages 39-45.

[8] C. Scott, "Performance Measures for Neyman-Pearson Classification ", *IEEE Transactions on Information Theory*, vol. 53, 2007,pp 2852-2863.

[9] W. Stoecker, K. Gupta, R. Stanley, R. Moss and B. Shrestha, "Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color", *Skin Research and Technology*, vol. 11, Number 3, August 2005 , pp. 179-184(6).

[10] K. Veropoulos, C. Campbell and N. Cristianini, "Controlling the Sensitivity of Support Vector Machines", In *Proceedings of IJCAI Workshop Support Vector Machines*, May, 1999.

[11] X. Yuan, N. Situ, G. Zouridakis, "A Narrow Band Graph Partitioning Method for Skin Lesion Segmentation", *Pattern Recognition*, accepted, under revision.