Cervical Cancer Detection Using SVM Based Feature Screening*

Jiayong Zhang and Yanxi Liu

The Robotics Institute, Carnegie Mellon University
{zhangjy,yanxi}@cs.cmu.edu

Abstract. We present a novel feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machines. It alleviates the "independence" assumption of traditional screening methods, e.g. those based on Information Gain and Augmented Variance Ratio, without sacrificing computational efficiency. We applied the proposed method to a bottom-up approach for automatic cervical cancer detection in multispectral microscopic thin PAP smear images. An initial set of around 4,000 multispectral texture features is effectively reduced to a computationally manageable size. The experimental results show significant improvements in pixel-level classification accuracy compared to traditional screening methods.

1 Introduction

Finding abnormal cells in PAP smear images (Fig. 1) is a "needle in a haystack" type of problem, which is tedious, labor-intensive and error-prone. It is therefore desirable to have an automatic screening tool such that human experts are only called for when complicated and subtle cases arise. Most researches on automatic cervical screening extract morphometric/photometric features at the cellular level in accordance with the "Bethesda System" rules [1]. However, accurate segmentations of cytoplasm and nucleus on cancer images are rather difficult due to the presence of blood, inflammatory cells, or thick cell clumps.

Using a micro-interferometric spectral imaging setup, we have obtained a set of multispectral Pap smear images. The wavelengths range from 400 nm to 690 nm, evenly divided into 52 bands. In [2], we propose a bottom-up approach to automatically detect cancerous regions in such images without the requirement of accurate segmentation. Our approach takes advantage of both the local multispectral and textural properties by learning a spatially-homogeneous discriminative filter. Cancerous regions are then detected from the filter output through a relatively simple procedure.

There are two critical issues that must be addressed in such a scheme: (1) what features should be extracted from multispectral images, and (2) how to remove irrelevant and/or redundant features from a pool of thousands of potential features to locate a feature subset that is well balanced between performance

* This research was supported in part by NIH award N01-CO-07119 and PA-DOH grant ME01-738.

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 873–880, 2004.
© Springer-Verlag Berlin Heidelberg 2004

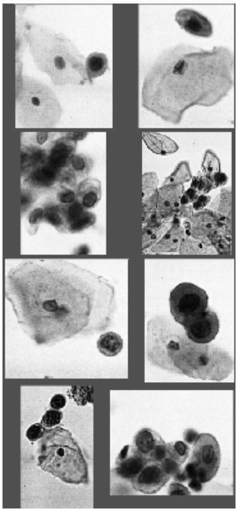

Fig. 1. Sample Pap smear images.

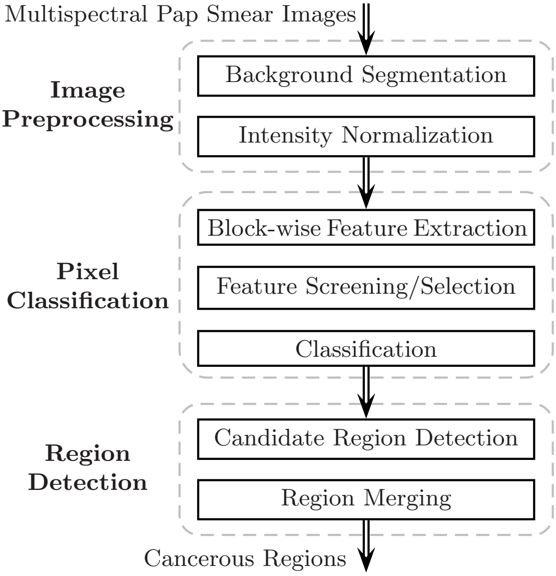

Fig. 2. Proposed bottom-up detection scheme.

and compactness. For the first issue, we have identified a feasible feature space of about 4,000 dimensions that well captures local multispectral and texture information. For the second issue, given that 4,000 dimensions is still intractable for traditional feature selection methods, we have employed two simple screening measures, i.e. Information Gain (IG) and Augmented Variance Ratio (AVR), to rule out irrelevant features. However, as each feature is evaluated independently, such screening methods may fail to capture all highly discriminative feature subsets, which are composed of individually less discriminative features.

In this paper, we present a novel feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machines [3]. The proposed relevance measures have several advantages: 1) As derived simultaneously for all dimensions, they do not only focus on single dimension as most existing measures do; 2) As the maximum margin boundary of SVM has been proven to be optimal in a structural risk minimization sense, they may better indicate the discriminative power of features; 3) As efficient routines for SVM training are available that can readily deal with huge number of features and samples, they do not sacrifice in computational cost. Our experimental results show significant improvements in pixel-level classification accuracy by using the proposed method.

2 Detection System Overview

A flowchart of the proposed bottom-up detection scheme is given in Figure 2. Here we summarize the main steps.

Preprocessing. We first segment cells from the background, which is much easier than traditional nuclei/cytoplasm segmentation. Then we normalize all band images by subtracting the spectral signature of the image background.

Feature Extraction. For each pixel, we extract various types of image features including: 1) Statistics of pixel intensity; 2) Daubechies 2 and Daubechies 16 asymmetric orthogonal wavelets; 3) Biorthogonal wavelet; and 4) Gabor wavelet. This procedure is applied to every band, resulting in a very high dimensional multispectral image feature set.

Feature Screening. The initial feature set is pruned by the proposed SVM based screening algorithm to remove those features irrelevant to the detection task. This is the main focus of the paper, and will be discussed in detail in Section 3. To further eliminate feature redundancy, Sequential Backward Selection (SBS) is applied to the surviving features of the screening procedure.

Discriminative Filtering. Let x be the final n-dimensional feature vector of a pixel after SBS selection. For each class (cancerous or normal), a modified version of quadratic discriminant is defined as

$$g(x) = \sum_{i=1}^{k} \frac{1}{\lambda_i} \left[\varphi_i^T(x-\mu)\right]^2 + \sum_{i=k+1}^{n} \frac{1}{\beta^2} \left[\varphi_i^T(x-\mu)\right]^2 + \ln\left[\beta^{2(n-k)} \prod_{i=1}^{k} \lambda_i\right] \quad (1)$$

where μ is the class mean, $\{\lambda_i, \varphi_i\}$ are the i-th eigenvalue and eigenvector of covariance matrix Σ, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and β is a positive constant. The discriminative filter output is then computed as $h(x) = g_{normal}(x) - g_{cancer}(x)$.

Region Detection. Based on the continuous output surface $h(x)$ from discriminative filtering, cancerous regions can be located by a relatively simple procedure (see Fig. 3 for example): 1) Smooth $H = \{h(x)\}$ with a Gaussian filter; 2) Find all local maxima m_i in H, and their corresponding effective regions R_i, defined as the points immediately around m_i with values above a fixed fraction (0.5) of $h(m_i)$; 3) For each R_i extract a geometric feature $G = C/L$, where C is the circumference of R_i, L is the distance from m_i to the boundary. Prune those R_i if $h(m_i) < 0.5 \max_i h(m_i)$ or $G < 2$, and generate the candidate region set; 4) Merge candidate regions that are overlapping.

3 SVM Based Feature Screening

Given a set of features in a classification problem, a basic question in many learning tasks is: what is the best feature subset for classification purpose? Although many feature subset selection methods have been proposed [4,5], few of them can be directly applied to domains with more than 100 dimensions. The huge feature dimension (near 4,000) and sample complexity (over 100,000) in our task make them computationally prohibitive. Alternatively, we present a new feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machine. Features are ranked according to these measures, and a subset is then selected via some statistical significance test.

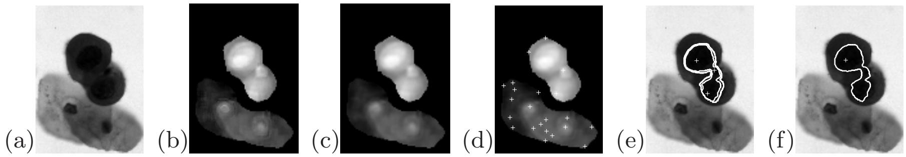

Fig. 3. An example of cancerous region detection. (a) Original image. (b) Scaled output surface from discriminative filtering. (c) Gaussian smoothing of (b). (d) Local maxima points found in (c). (e) Contours of candidate cancerous regions. (f) Merged result.

The SVM decision function of a two-class problem can be written as

$$h(x) = w \cdot \Phi(x) + b = \sum\nolimits_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \tag{2}$$

where $x_i \in \mathbb{R}^d$ is the training sample, and $y_i \in \{\pm 1\}$ is the class label of x_i. A transformation $\Phi(\cdot)$ maps the data points x of the input space $\mathbb{R}^d$ into a higher dimensional feature space $\mathbb{R}^D, (D \geq d)$. The mapping is performed by a kernel function $K(\cdot, \cdot)$ which defines an inner product in $\mathbb{R}^D$. The parameters $\alpha_i \geq 0$ are optimized by finding the hyperplane in feature space with maximum distance to the closest image $\Phi(x_i)$ from the training set, which reduces to solving a linearly constrained convex quadratic program. In the general case of nonlinear mapping Φ, SVM generates a nonlinear boundary $h(x) = 0$ in the input space.

Given any two points $z_1, z_2 \in \mathbb{R}^d$ such that $h(z_1)\, h(z_2) < 0$, a surface point $s = \alpha z_1 + (1 - \alpha) z_2, \alpha \in [0, 1]$, can be found by solving the following equation with respect to α:

$$h(s) = h\left(\alpha z_1 + (1 - \alpha) z_2\right) = 0 \tag{3}$$

The unit normal vector $N(s)$ at the boundary point s is then given by

$$N(s) = \nabla h(s) / \|\nabla h(s)\| \tag{4}$$

where $\nabla h(s) = \partial h(s) / \partial s = \sum_{i=1}^{n} \alpha_i y_i \, \partial K(s, x_i) / \partial s$. $N(s)$ identifies the orientation in the input space along which the projected training data are well separated locally around the neighborhood of s. Therefore, the orientation difference between $N(s)$ and any direction u can be used to measure the local discriminative relevance for that direction at s. Formally, we measure this difference by $|u^T N(s)|$, or equivalently $u^T N(s) N(s)^T u$. To summarize all the local feature relevance information, we compute the decision boundary scatter matrix as

$$M = \int_{\mathcal{B}} N(s) N^T(s) p(s) \, ds \tag{5}$$

and a global relevance measure for direction u as $u^T M u$. When sample-size is finite, M can be replaced by the sample estimate $\hat{M} = \sum_{i=1}^{l} \hat{N}(\hat{s}_i)\, \hat{N}(\hat{s}_i)^T / l$, where $\hat{s}_i$ are l points sampled from the estimated decision boundary. This global relevance measure can be readily extended to multi-category problems by repeating the procedure in either one-vs-all or pairwise mode. Now we summarize the SVM based feature screening algorithm as follows.

```
Input: $n$ sample pairs $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{k\}_{k=1}^{Q}$.
Output: $d$ nested feature subsets $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{S}_d$ such that $\dim(\mathcal{S}_m) = m$.
Algorithm:
S1 For $k = 1$ to $Q$
S2 Divide the $n$ samples into two subsets $T^+ = \{x_i | y_i = k\}$ and $T^- = \{x_i | y_i \neq k\}$. Learn a SVM decision function $h(x)$ using $T^+$ and $T^-$.
S3 Sort the $n$ samples in ascending order by the absolute function output values $|h(x_i)|$. Denote the subset consisting of the first $r$ samples as $T'$.
S4 Select $l$ pairs of points $\{(z_1^j, z_2^j)\}_{j=1}^{l}$ from $T'$ randomly such that $h(z_1^j)\,h(z_2^j) < 0$. For each pair solve equation (3) to an accuracy of $\epsilon$, and thus get $l$ estimated boundary points $\{\hat{s}_j\}_{j=1}^{l}$.
S5 Compute the unit surface norm $\hat{N}(\hat{s}_j)$ at $\hat{s}_j$ according to equation (4), and estimate the decision boundary scatter matrix as $\hat{M}_k = \sum_{j=1}^{l} \hat{N}(\hat{s}_j)\hat{N}(\hat{s}_j)^T$.
S6 End (For $k = 1$ to $Q$)
S7 Compute $\hat{M} = \sum_{k=1}^{Q} \hat{M}_k / Q$, and denote its diagonal value as $\{\hat{\lambda}_j\}_{j=1}^{d}$.
S8 Sort feature directions $\{u_j\}_{j=1}^{d}$ descendingly by $\{\hat{\lambda}_i\}_{j=1}^{d}$. Let $\mathcal{S}_m = \{u_j^{sort}\}_{j=1}^{m}$.
```

Note that first, we prune those training samples far away from the decision boundary in locating the boundary points. This helps to reduce computational cost and suppress the negative influence of outliers. Second, we adopt the one-vs-all approach for solving Q-class problems with SVMs. Totally Q classifiers need to be trained, each of which separates a single class from all remaining classes. Third, the complexity of the algorithm can be controlled by several parameters including l, the number of boundary points to be sampled, and ϵ, the accuracy of the root to equation (3). Our experience suggests that the algorithm is not very sensitive to the choice of these parameters. Finally, we have used p-degree polynomial kernels in our experiments, where $p = 2$.

It can be proven that a feature u is irrelevant if and only if $u^T M u$ equals zero. In theory we can prune all irrelevant features via this screening method. However, inherent uncertainty in our estimation prevent us from doing so. A more practical reason is that, features' contribution to discrimination may be unevenly distributed that the subset dimension can be significantly reduced while achieving *almost* the same accuracy. Therefore model selection technique is required in order to decide an appropriate subset. This problem will not be discussed in this paper, but we want to point out that nested subsets generated by SVM based screening can easily facilitate such explorations.

4 Experiments and Analysis

We evaluated the proposed SVM based screening algorithm and the resultant cervical cancer detection system on a multispectral PAP smear image database containing 40 images (each has 52 spectral bands) with a total of 149 cells (41 cancerous and 108 normal). The image size ranges from 93x64 to 300x227. First, all images are preprocessed to remove the background and normalize intensity by setting background spectral signature to zero. Then for each pixel to be

Table 1. Various dimensions before and after SVM based feature screening.

	DB2	DB16	Bio2.2	Gabor	Combined
Original Dim.	800	800	900	1200	3700
After Screening	48	42	52	30	68

classified, various image features are extracted in a 16 × 16 block (the block size is chosen via cross-validation) around it in each band, as described in section 2. Thus a very high dimensional multispectral texture feature vector is associated with each pixel. At pixel level, we collect a total of 156,732 sample vectors (26,064 positive and 132,063 negative) from all 40 images. As samples from the same image are often highly correlated, they are always kept as a whole when partitioning training and test sets in all the following experiments.

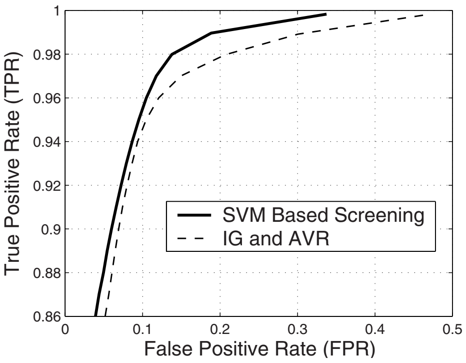

Fig. 4. Comparison between SVM and IG+AVR based screenings.

We evaluated the selected 68 features on the original full sample set using the modified quadratic discriminant. The ROC curve is plotted in Figure 4 against the ROC curve of IG+AVR screening depicted for comparison. It is easy to observe that SVM based screening outperforms IG+AVR, especially when the True Positive Rate (TPR) is high.

Pixel Classification. We investigated the effect on pixel-level classification by replacing IG and AVR feature screening [2] with the proposed method. In order to reduce the training complexity, a total of 29,487 samples were randomly selected (13,022 positive and 16,465 negative). SVM based feature screening method with p-degree polynomial kernels was applied to each of four types of wavelet features respectively. For each type of wavelet, images were randomly divided into training set (32 images) and test set (8 images) for a number of times. Each time we record False Positive Rate (FPR) on the test set versus subset dimension. Then we averaged these FPR curves, based on which a proper feature dimension m was manually selected. After that we collected all features ever appeared among the top m features in each image partition, and regarded them as the selected features for that wavelet type. Then we put together all the selected features from four types of wavelets, and applied the SVM based feature screening algorithm. Again we did random partition of training and test images, and selected a proper dimension m' based on the average FPR curve. Various dimensions before and after feature screening are summarized in Table 1.

Finally, we applied Sequential Backward Selection (SBS) to the 68 survivals of SVM based screening to investigate their redundancy. 8-fold cross validation error on the training set was chosen as the evaluation function in SBS. The aver-

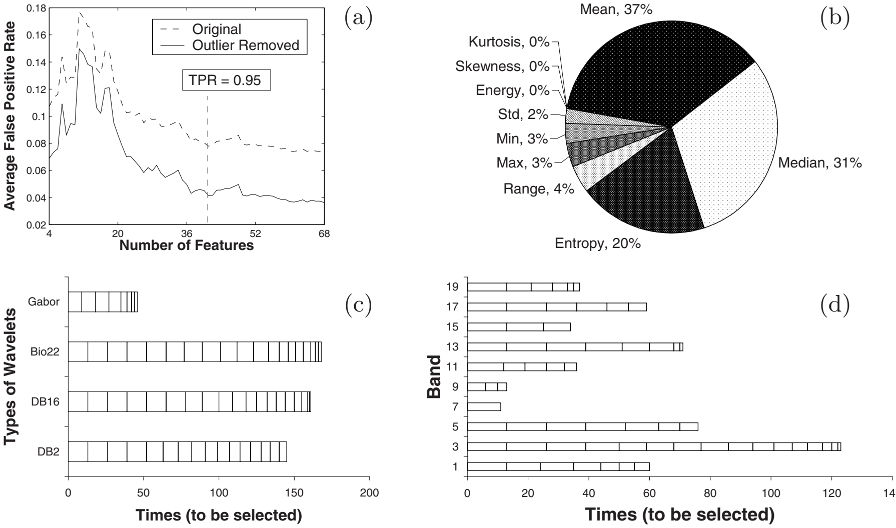

Fig. 5. (a) Average FPR versus subset dimension in sequential backward selection. Analysis of the selected feature subsets with respect to their feature type and spectral band distribution is also provided for us to gain some insight of the selected features. Plots shown are frequency histograms of selected statistics (b), wavelet features (c), and spectral bands (d). Each short segment in (c-d) corresponds to a particular feature.

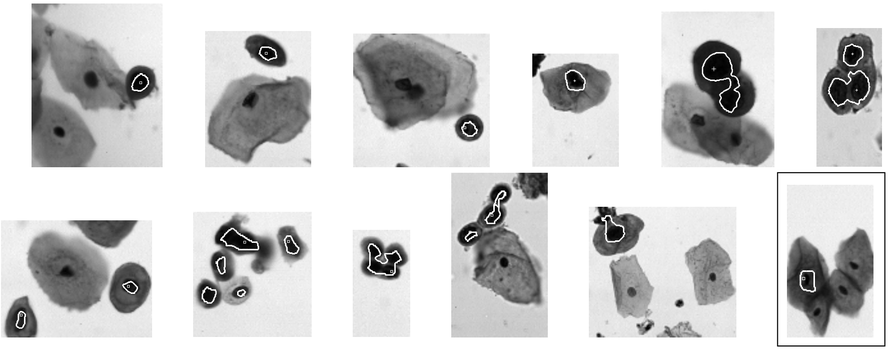

Fig. 6. Example cancerous cell detection results. All images contain one or more cancerous cells except the last one (with frame), which is a false positive case.

age accuracy over 13 test runs is depicted in Fig 5(a). It is observed that feature dimension can be consistently reduced below 40 with little loss of accuracy. We analyzed those features that rank among the top 40 in any of the 13 runs with respect to their feature type and spectral band distribution, and the results are summarized in Figure 5(b-d). Note that distributions of discriminative features are not uniform. For instance, 86.9% features are from 3 out of 10 types of statistics (mean 36.5%, median 30.8%, entropy 19.6%). Over 15 features are selected from spectral band 3 while only 1 from band 7.

Region Detection. As the number of available images is small, we evaluate the performance of the complete detection system by leave-one-out cross validation method. Each time 39 images are used to train the pixel classifier, and one image is reserved for test. Some typical detection results are shown in Figure 6. Among the 149 cells distributed in 40 images, one cancerous cell is missed (TPR $= 40/41 \approx 98\%$), and one normal cell is falsely detected (FPR $= 1/108 \approx 1\%$).

5 Related Work and Conclusion

In this paper, we presented a novel SVM-based feature screening method. Guyon et al. [6] proposed a feature ranking scheme by linear SVMs. The basic idea is to use the magnitude of the weights of a linear discriminant classifier as an indicator of feature relevance. Our method can be considered as a nonlinear extension of this linear scheme. SVM boundary has also been used in locally adaptive metric techniques to improve k-NN performance [7]. Measures of local feature relevance are computed by the surface normal near the query, from which a local full-rank transformation is derived. Such local methods need to perform k-NN procedure multiple times in the original high-dimensional space. On the contrary, our method tries to globally characterize the discriminative information embedded in the SVM decision boundary. It generates global feature relevance measures, and thus is computationally more efficient.

We applied the proposed method to multispectral Pap smear image classification for cervical cancer detection. Comparative experiments show significant improvements on pixel-level classification accuracy using the new feature screening method. We have shown the effectiveness of image feature screening/selection in cancerous cell detection on a novel image modality (multispectral image). A much larger PAP smear image set and an even richer image feature space will be used to further validate our method.

References

1. Kurman, R., Solomon, D.: The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses. Springer-Verlag, New York (1994)
2. Liu, Y., Zhao, T., Zhang, J.: Learning multispectral texture features for cervical cancer detection. In: Proc. Int. Symp. Biomedical Imaging. (2002) 169–172
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK (2000)
4. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence **97** (1997) 245–271
5. John, G., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Proc. 11th Int. Conf. Machine Learning. (1994) 121–129
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
7. Domeniconi, C., Gunopulos, D.: Adaptive nearest neighbor classification using support vector machines. In: Advances in NIPS 14, MIT Press (2001)