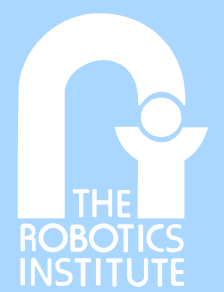# Cervical Cancer Detection Using SVM Based Feature Screening

Jiayong Zhang & Yanxi Liu, The Robotics Institute, Carnegie Mellon University

Introduction

Annually, over 50 million Pap smears are done in US and over 60 million in the rest of the world. Finding abnormal cells in Pap smear images remains to be a "needle in a hay-stack" type of problem. Highly accurate, automated screening systems are in great need.

Previous works mostly extract shape features at the cellular level in accordance with the "Bethesda System" rules. However, due to image segmentation errors, cellular shape analysis can be rather difficult.

We investigate this problem on a novel image modality (multispectral), and propose a bottom-up approach to automatically detect cancerous regions without the requirement of accurate segmentation.
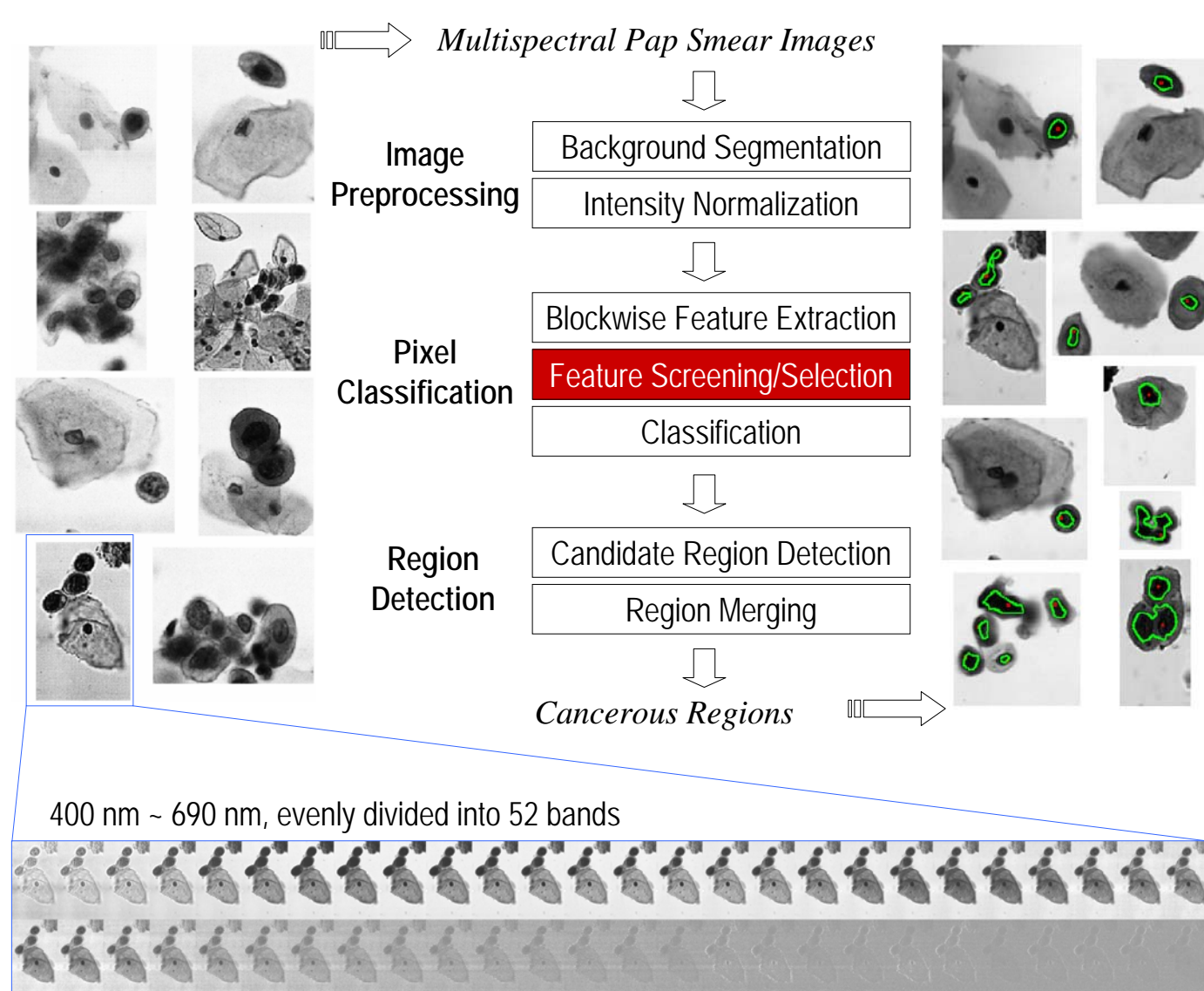
By exploring an initial image feature space of nearly 4,000 dimensions that captures local multispectral and texture information, we found that existing feature subset selection algorithms are computationally challenged by such large sized feature set.

One alternative is to use simple feature screening measures, e.g. Information Gain (IG) and Augmented Variance Ratio (AVR), to rule out irrelevant features. However, by evaluating each feature independently, they may fail to capture all highly discriminative subsets, which could be composed of individually less discriminative features.

In this work, we present a novel feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machines. Advantages:
• Relevance measures (feature weights) derived simultaneously for all dimensions
• Optimal in Structural Risk Minimization sense ⇨ Better discriminative power indicator
• Efficient SVM training ⇨ Little sacrifice in computational cost

Detection System Overview


Multispectral Pap Smear Images

Image Preprocessing
- Background Segmentation
- Intensity Normalization

Pixel Classification
- Blockwise Feature Extraction
- Feature Screening/Selection
- Classification

Region Detection
- Candidate Region Detection
- Region Merging

Cancerous Regions

400 nm ~ 690 nm, evenly divided into 52 bands

Multispectral Texture Features

• Statistics (10): maximum, minimum, range, median, mean, standard deviation, energy, skewness, kurtosis and entropy.
• Wavelets (4): DB2 and DB16 (Orthogonal), Bior2.2 (Bi-orthogonal), Gabor (Non-orthogonal).

These features are generated per pixel, per spectral band.

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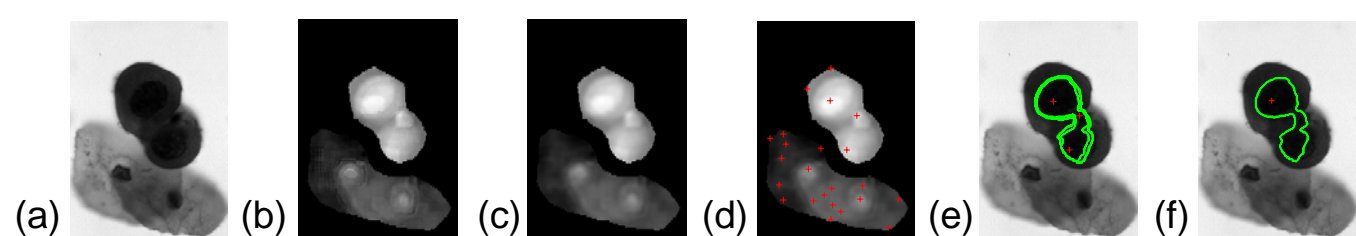
✍ An example of cancerous region detection.
(a) Original image. (b) Scaled output surface from discriminative filtering.
(c) Gaussian smoothing of (b). (d) Local maxima points found in (c).
(e) Contours of candidate cancerous regions. (f) Merged result.

Feature Screening

Concept: A greedy feature selection method. Rank features and discard those whose ranking criterions are below the threshold.

Problem: What is a good ranking criterion (relevance measure or feature weight)?

Intuition: *Large feature weight if data are well separated along that feature direction*

Observations:
• Decision boundary $h(s)$ encodes all discriminative information.
• $h(s)$ of SVM has an analytical form.
$$h(\boldsymbol{x}) = \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) + b = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b = 0$$
• Boundary normal $N(s) = \nabla h(s) / \|\nabla h(s)\|$ identifies the direction along which the data are locally well separated around the neighborhood of boundary point s.

Conclusions:
• Given any direction u, a local relevance measure can be defined as the "consistency" between $N(s)$ and u (e.g. $|u^T N(s)|$, $u^T N(s) N(s)^T u$).
• Decision Boundary Scatter Matrix (DBSM) summarizes local discriminative directions over the whole decision boundary.
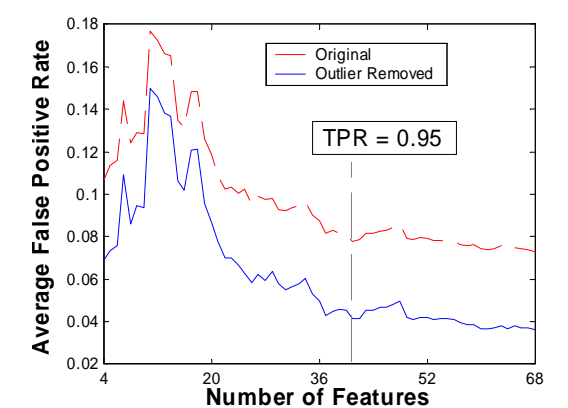$$M = \int_B N(s) N^T(s) p(s) ds$$
• Given any direction u, a global relevance measure can be defined as the "consistency" between M and u (e.g. $u^T M u$).
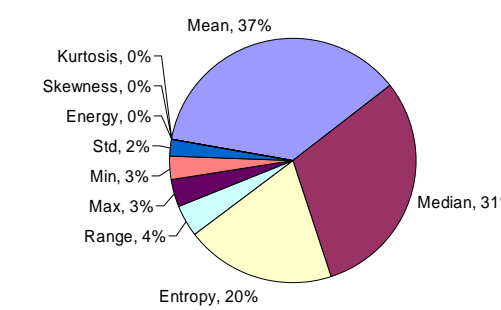
Evaluation

	DB2	DB16	Bior2.2	Gabor	Combined
Original Dimensions	800	800	900	1200	3700
After IG Screening	74	76	71	51	144
After AVR Screening	47	35	41	21	45
After SVM Screening	48	42	52	30	68

☞ Applying sequential backward selection to surviving features of screening procedure leads to further reduction in subset sizes.

✎ Various dimensions before and after feature screening.

Mean, 37%
Kurtosis, 0%
Skewness, 0%
Energy, 0%
Std, 2%
Min, 3%
Max, 3%
Range, 4%
Median, 31%
Entropy, 20%

☞✎ Analysis of the selected feature subsets with respect to their feature type and spectral band distribution provides some insights into the interpretations of the results.
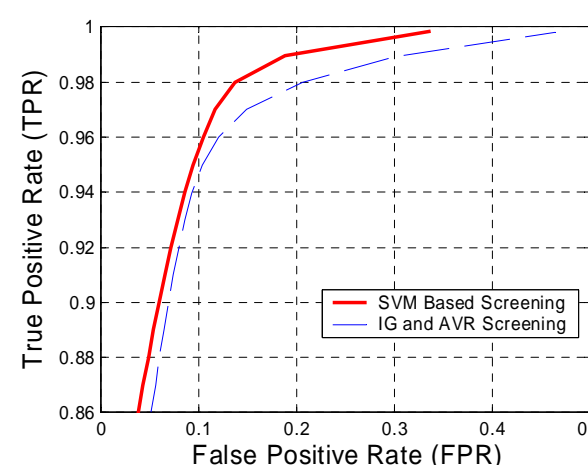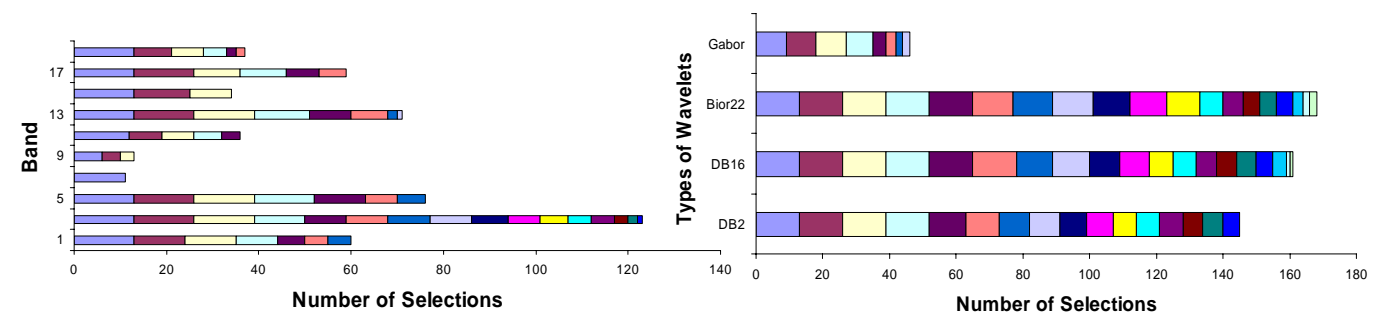
✎ Pixel-level classification. Comparison between SVM and IG+AVR screenings.

✎ Region-level detection. Leave-one-out system evaluation.

	CR	TPR	FPR
# cells	147/149	40/41	1/108
ratio %	98.7	97.6	0.09

Conclusion

We show the effectiveness of image feature screening/selection in cancerous cell detection on a novel image modality (multispectral). An initial set of around 4,000 multispectral texture features is effectively reduced to a computationally manageable size. Comparative experiments show significant improvements on pixel-level classification accuracy using the new feature screening method. A much larger PAP smear image set and an even richer image feature space will be used to further validate our method.

Acknowledgments