Automatic Segmentation and Supervised Learning Based Selection of Nuclei in Cancer Tissue Images – Supplementary Material

Kaustav Nandy[1], Prabhakar R. Gudla[1], Ryan Amundsen[2], Karen J. Meaburn[3], Tom Misteli[3] and Stephen J. Lockett[1]

[1]*Optical Microscopy and Analysis Laboratory, Advanced Technology Program, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, 21702 USA*

[2]*Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI, 48109 USA*

[3]*Cell Biology of Genomes, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892 USA*

Running headline: Segmentation and selection of tissue nuclei

Corresponding Author: Kaustav Nandy

Address: Optical Microscopy and Analysis Laboratory, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Post Office Box B, Frederick, MD 21702

Phone: +1 3018466109, Fax: +1 3018466552

Email : nandyk@mail.nih.gov

Tree Based Merging Algorithm

Fig. 1 shows the detailed steps of the tree based merging algorithm used to merge the watershed segmentation fragments using an elliptical shape modelling of cell nuclei.

Feature Set

The 64 dimensional feature set used to characterize segmented objects for classification is provided below.

1. CCBendingEnergy
2. Feret (3 dimensions)
3. GreyInertia (2 dimensions)
4. GreyMu (3 dimensions)
5. Inertia (2 dimensions)
6. Mass
7. Mean object intensity
8. Mu (3 dimensions)
9. P2A
10. Perimeter
11. PodczeckShapes (5 dimensions)
12. Size
13. StdDev
14. DimensionsEllipsoid - major axes
15. DimensionsEllipsoid - minor axes
16. DimensionsEllipsoid - eccentricity
17. DimensionsEllipsoid - minor/major ratio
18. MajorAxes (4 dimensions)
19. ConvexRatio - Object Area/Convex Hull Area

20. ConvexRatio - Convex Hull Area - Object Area
21. RadiusStats - entropy of all radii
22. RadiusStats - range of all radii
23. RadiusStats - variance of all radii
24. GradientStats - magnitude sum
25. GradientStats - mean
26. GradientStats - range
27. GradientStats - variance
28. Autocorrelation
29. Contrast
30. Correlation (2 dimensions)
31. Cluster Prominence
32. Cluster Shade
33. Dissimilarity
34. Energy
35. Entropy
36. Homogeneity (2 dimensions)
37. Maximum probability
38. Sum of sqaures: Variance
39. Sum average
40. Sum variance
41. Sum entropy
42. Difference variance
43. Difference entropy

44. Information measure of correlation1
45. Information measure of correlation2
46. Inverse difference (INV)
47. Inverse difference normalized (INN)

Normalization

Table 1 shows the feature normalization methods used, namely linear scaling to unit range, Z-score based normalization, linear scaling to unit variance, transformation to uniform distribution and rank normalization.

Feature Dependency Ranking

Fig. 2 shows the dependency ranking values for the 64 features with respect to the output object labels. Some of the features that had a very high dependency ranking with respect to the output labels were P2A (feature 17 in the figure), the first 4 PodczeckShapes (features 19 to 22 in the figure), ConvexRatio - Object Area/Convex Hull Area (feature 34 in the figure) and RadiusStats - range of all radii (feature 37 in the figure).

3D Feature Plot

The 3 features having the highest variances after principal component analysis (PCA) for the training set segmented objects are plotted in 3D in Fig. 3. The figure shows that although the well segmented nuclei do form a cluster in the 3D space, due to considerable overlap with the rest of the objects a non linear classifier is required for the application.

Objects Selected by the PRE

Fig. 4 shows five sample original nuclei channel images (a,c,e,g,i) along with their segmentation outputs (b,d,f,h,j). The objects with green boundaries were selected by the PRE as well-segmented and those with red boundaries were rejected.

Boundary accuracy parameters

Fig. 5 shows the plots of the 3 boundary accuracy parameters for 20 synthetic nuclei segmented manually (red plot) and using 2D dynamic programming (DP) based method (blue plot). The plot reveals that when

compared to the control nuclei mask, the 2D DP segmentation shows better segmentation accuracy in general compared to the manual segmentation.

FISH copy number distribution

Table II shows the fluorescence *in situ* hybridization (FISH) signal copy number ditribution for normal, cancer and non-cancerous breast disease tissue sections.

Acknowledgment

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research and by a Department of Defense Breast Cancer Idea Award to Tom Misteli.

The Office of Human Subjects Research (OHSR) at the National Institutes of Health, USA determined on January 2 2008 that federal regulations for the protection of human subjects do not apply to this research project. The human material used in this study had been de-identified before any of the authors received it.

Fluorescence imaging was performed at the National Cancer Institute Fluorescent Imaging Facility, Bethesda, MD, USA.

List of Figures

Tree Based Merging Algorithm

```
1   For each cluster of Watershed fragments
2       Create Region Adjacency Graph (RAG)
3       For each node in the RAG
4           Create merge tree and calculate merge criteria for each merged object using optimal
            ellipse fitting and calculate overlap area
5               For each merged object
6                   If merged object and optimal ellipse has overlap area  > 80%
7                       Merge Objects
8                       Go to 1
9                   Else
10                      Go to 5
11                  End
12              End
13      End
15  End
```

Figure 1: Tree based merging algorithm

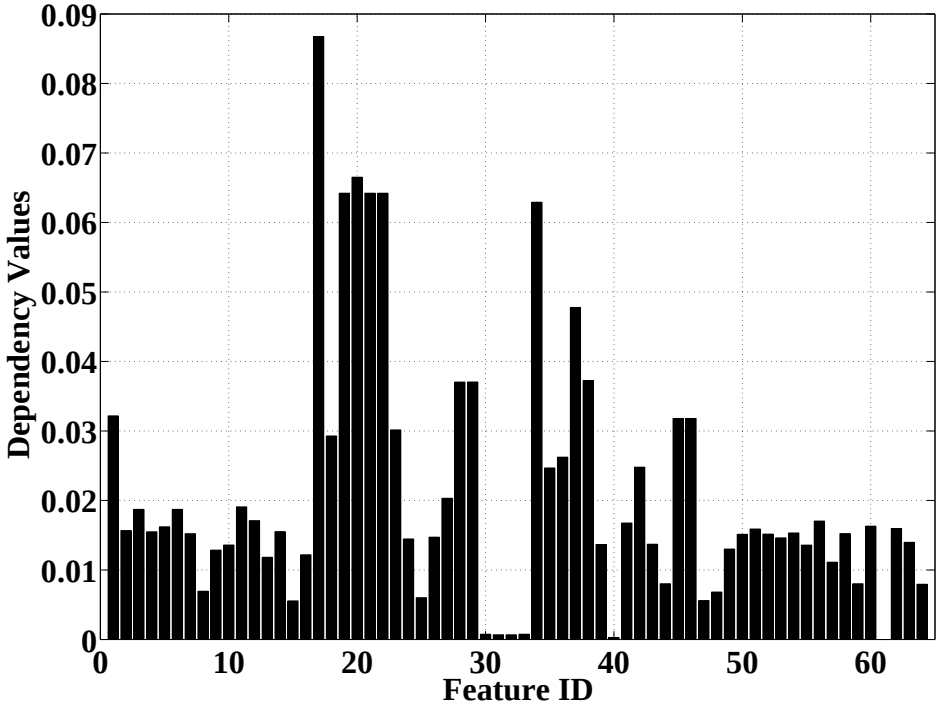

Figure 2: Dependency ranking of 64 features

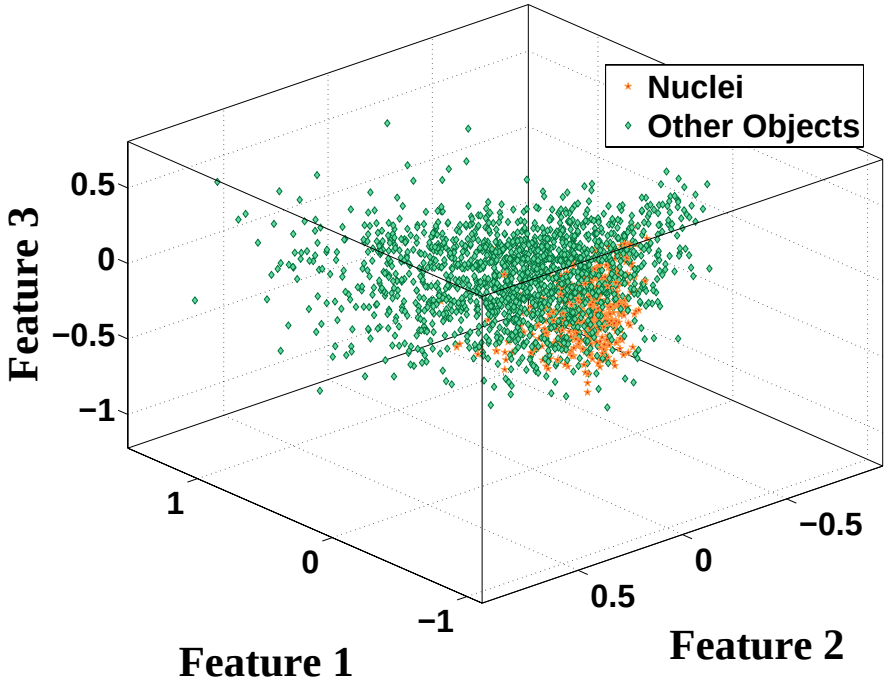

Figure 3: 3 dimensional plot of the top 3 features after PCA

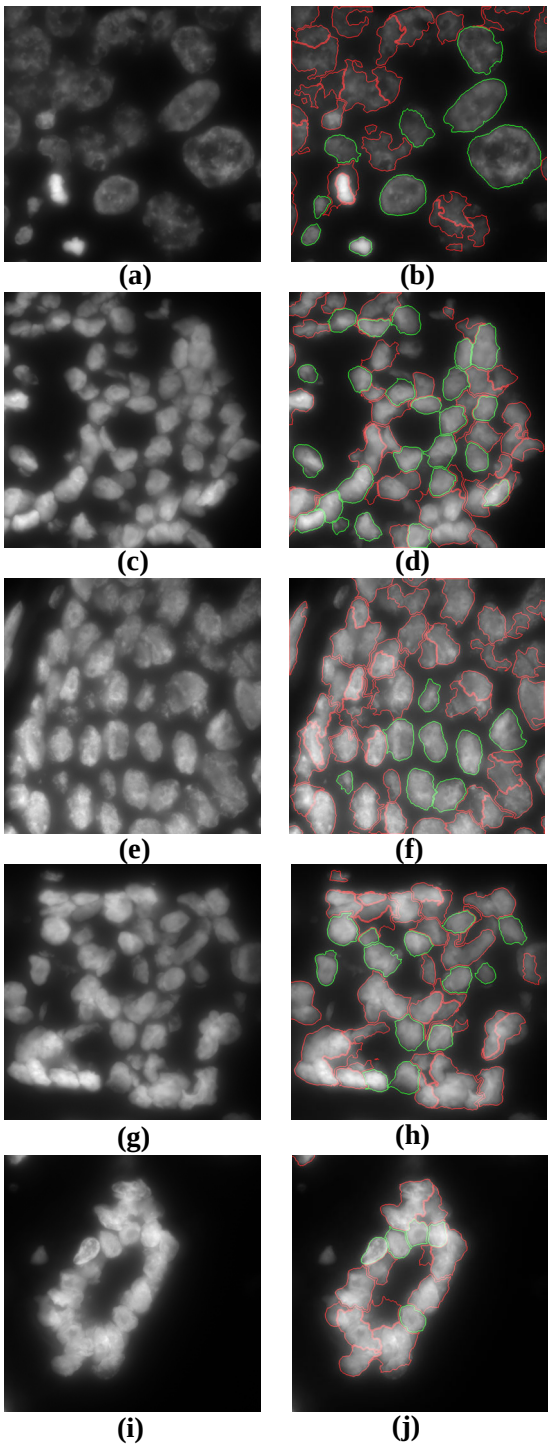

Figure 4: Samples of objects selected by the PRE as well-segmented. (a,c,e,g,i) show 5 sample original image nuclei channels and (b,d,f,h,j) are the corresponding segmentation outputs. Objects with green boundaries were selected by the PRE as well-segmented and those with red boundaries were rejected.

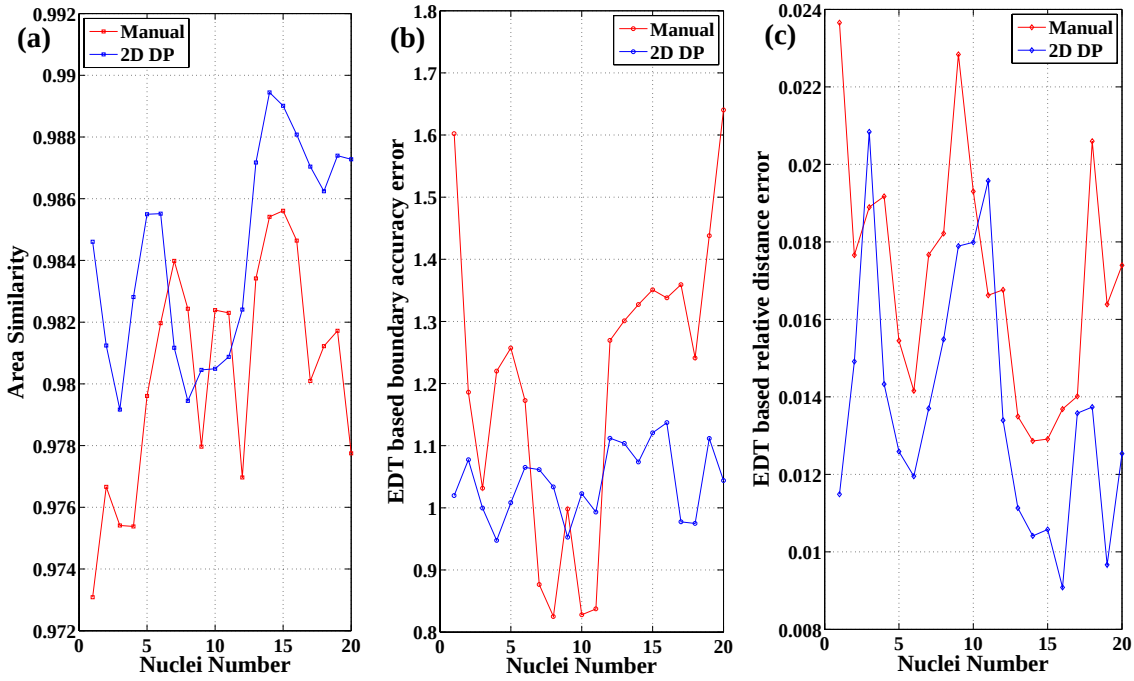

Figure 5: Plots showing (a) area similarity, (b) mean EDT based boundary error per pixel and (c) difference in EDT based relative internal distance measure for 20 synthetic nuclei segmented manually and using 2D DP

List of Tables

Table 1: Normalization techniques. x_i is the i^{th} feature vector

	Normalization Method	Procedure
1	Linear Scaling to Unit Range l = lower bound u = upper bound	$\tilde{x}_i = \frac{x_i - l}{u - l}$
2	Z-Score μ = Mean σ = Standard Deviation	$\tilde{x}_i = \frac{x_i - \mu}{\sigma}$
3	Linear Scaling to unit variance μ = Mean σ = Standard Deviation	$\tilde{x}_i = \frac{\frac{x_i - \mu}{3\sigma} + 1}{2}$
4	Transformation to Uniform Distribution	$\tilde{x}_i = F_{x_i}(x_i)$
5	Rank Normalization $\tilde{x}_{ij}$ = Normalized location of the j_{th} element of feature vector σ = Standard Deviation	$\tilde{x}_{ij} = \frac{rank_{x_{i1},\ldots,x_{iN}}(x_{ij}) - 1}{N - 1}$

Table 2: Details of FISH analysis showing number of nuclei analyzed, number of FISH spots analyzed and gene copy number distribution for both manual(M) and automatic(A) analysis

Datasets	Number of Nuclei		Total number of red FISH signals		Number of Nuclei with Gene copy number									
					0		1		2		3		>3	
	M	A	M	A	M	A	M	A	M	A	M	A	M	A
N1-N4	536	670	943	1191	10	106	115	160	400	287	8	66	1	51
C1-C14	1965	2387	2542	2952	306	761	913	828	597	506	100	168	31	124
B1-B5	699	673	1010	833	72	196	306	227	301	180	30	47	3	23