Automatic Content Analysis of Uterine Cervix Images using Computerized Tools

By

Shiri Gordon

THESIS SUBMITTED FOR THE DEGREE OF "DOCTOR OF PHILOSOPHY"
SUBMITTED TO THE SENATE OF TEL-AVIV UNIVERSITY

July 2008
Automatic Content Analysis of Uterine Cervix Images using Computerized Tools

By

Shiri Gordon

THESIS SUBMITTED FOR THE DEGREE OF “DOCTOR OF PHILOSOPHY” SUBMITTED TO THE SENATE OF TEL-AVIV UNIVERSITY

This Research Work was Carried Out at Tel-Aviv University The Faculty of Engineering

Under the Supervision of Dr. Hayit Greenspan

July 2008
This work was carried out under the supervision of

Dr. Hayit Greenspan
Acknowledgements

I wish to express my deepest gratitude to my advisor Dr. Hayit Greenspan who introduced me to the world of medical image processing and analysis. For her companionship, great support, constructive ideas and guidance throughout my PhD studies, thank you.

I would like to thank the present and the former members of the medical image processing laboratory at Tel-Aviv University for their friendship and encouragement. Special thanks to Gali Zimmerman, Hila Dvir and Shelly Lotenberg with whom I had the pleasure to collaborate and whose devotion and ideas contributed so much to this research. Because of you our laboratory was always a fun place to be in.

I would like to thank Mr. Rodney Long and Dr. Sammer Antani from the National Library of Medicine, National Institute of Health and to Dr. Jose Jeronimo from the National Cancer Institute, National Institute of Health, for enabling me to explore the cervigram database and to conduct this research. Thank you for your time, support, ideas and encouragement that motivated me throughout the research.

I would also like to thank my friends and colleagues - Sigal Trattner, Tammy Riklin-Raviv, Leah Bar and Chen Sagiv for their friendship and fruitful conversations. Special thanks to Dr. Jacob Goldberger who introduced spectral clustering to me and to Dr. Nir Sochen, who contributed to my knowledge on calculus of variations.

I want to thank my friends from the department of Biomedical Engineering at Tel-Aviv University for friendly and supportive atmosphere.

I wish to express my sincere gratitude to my parents Nurit and Dalik Gordon and to my mother-in-law Silvy Havatzelet, for being there during the tough moments, enabling
me to finish this work.

Finally, to my dear husband and my best friend, Dan, thank you for your love and support. To you, and to my sweet daughters Amit and Yuval, all the love in the world.
Abstract

This research is focused on automatic analysis of optical images of the uterine cervix, termed “cervigrams”. Cervigrams are currently being investigated as a means for detection, diagnosis and basic research of cervical cancer. The National Cancer Institute (NCI), National Institute of Health (NIH), has collected a unique medical repository of digital cervicographic images in longitudinal multi-year studies. NCI, together with the National Library of Medicine (NLM), NIH, is developing a unique Web-accessible database of the digitized cervix images to study the evolution of lesions related to cervical cancer. The repository contains a large amount of cervigrams with diverse cervix content and varying illumination conditions.

The cervigrams within the NIH archive have no attached annotation that describes their visual features. Important information that relates to the visual content of the cervigrams is currently inaccessible and the automated analysis of the cervigrams by computerized tools is highly desirable. This thesis presents a multi-stage framework for segmenting and labeling regions of medical and anatomical interest within the cervigrams. In particular the work focuses on the following tasks for which specific tools were developed: 1) Extraction of the cervix region and fine detection of the cervix boundary; 2) Illumination correction and intensity normalization; 3) Cervix tissues segmentation.

Cervix boundary extraction is done using an active contour framework which is based on local convexity features, color and shape. The illumination correction step is based on a generalized expectation-maximization algorithm that interleaves pixel classification with estimation of class distribution and illumination field parameters.
For cross-image analysis a normalization of the image dynamic range is conducted using prior knowledge on cervix tissues intensity distribution. Two unsupervised tissue segmentation frameworks are examined next. The first framework is based on a probabilistic pixel-based segmentation algorithm. In this framework, Gaussian models are learned for tissues of interest within the cervix in a supervised manner. Utilizing the tissue models and a maximum a posteriori version of the expectation-maximization algorithm, a full range of models is enabled, from models learned in a totally unsupervised manner to models learned in a supervised manner. These models are used for probabilistic segmentation. The second framework presents a new methodology that enables the segmentation of elongated, thin and non-convex regions within the cervix. The framework transitions from pixels to a set of small coherent regions (superpixels), which are grouped into larger, perceptually similar regions. This is done utilizing a new graph-cut criterion and different variations of the agglomerative clustering algorithm. Superpixels similarity in this framework is computed via a combined region and boundary information measure.

The research includes a thorough validation of the automated analysis framework using several image sets of cervigrams that were manually labeled by NCI experts. The last part of the thesis includes a special chapter that addresses the task of segmentation evaluation when the markings of multiple experts are available. This chapter presents some general concepts, with a focus on the cervigrams archive.

The current research is one of the first attempts to automatically analyze the cervigrams within the NIH archive. It provides a major step forward in answering the defined tasks and in identifying the overall challenges of a complete cervigram analysis system.
Contents

List of Abbreviations & Symbols ix

List of Figures xvi

List of Tables xix

1 Introduction 1

1.1 Cervicography in cervical cancer research . . . . . . . . . . . . . . . . . 1
1.2 Automated cervigram analysis challenges . . . . . . . . . . . . . . . . . 4
1.3 Main objectives and significance . . . . . . . . . . . . . . . . . . . . . . 7
1.4 Dissertation overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2 Preprocessing and cervix boundary detection 13

2.1 Extracting ROI and removing specular reflection artifacts . . . . . . . 13
  2.1.1 Initial automated ROI detection . . . . . . . . . . . . . . . . . 13
  2.1.2 Detection and elimination of specular reflection . . . . . . . . . 15
2.2 Cervix boundary detection . . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.2.1 Curve evolution based on image data . . . . . . . . . . . . . . . 18
  2.2.2 Curve evolution based on a prior shape model . . . . . . . . . . 24
  2.2.3 Combination of the data and shape terms in the curve evolution
       process . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
2.3 Illumination correction and intensity normalization . . . . . . . . . . . 31
  2.3.1 Intensity normalization . . . . . . . . . . . . . . . . . . . . . . . 37
3 Cervix tissues segmentation 39

3.1 Pixel-based segmentation of the cervix tissues 40
   3.1.1 Cervix tissue segmentation using a Gaussian Mixture Model 41
   3.1.2 Learning the tissue models 42
   3.1.3 Incorporation of context-based rules 44

3.2 Cervix tissues segmentation via clustering of superpixels 46
   3.2.1 Superpixels generation 48
   3.2.2 Region & edge similarity matrix 50
   3.2.3 Agglomerative clustering of superpixels 53

4 Experimental results 57

4.1 Cervix boundary detection results 59
4.2 Illumination correction results 67
4.3 Analysis of probabilistic pixel-based cervix tissues segmentation 71
4.4 Analysis of segmentation via clustering of superpixels 75
   4.4.1 Quality analysis of superpixels overlap with cervix tissues 77
   4.4.2 Quality analysis of the similarity matrix 78
   4.4.3 Quality analysis of the agglomerative clustering framework 83
4.5 Preliminary tissue classification results 101

5 Evaluation of multiple-experts markings 108

5.1 Introduction to multi expert evaluation 108
5.2 The STAPLE algorithm 111
5.3 Performance analysis based on multi-expert ground truth maps 116
   5.3.1 Measuring segmentation complexity 116
   5.3.2 Evaluating automatic segmentation results 118
5.4 Experiments and Results 121
   5.4.1 Evaluation of the segmentation complexity descriptors 122
   5.4.2 A comparison between the complexity of the acetowhite and cervix boundary segmentation tasks 127
## List of Abbreviations & Symbols

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIN</td>
<td>Cervical Intraepithelial Neoplasia</td>
</tr>
<tr>
<td>HPV</td>
<td>Human Papillomavirus</td>
</tr>
<tr>
<td>Pap</td>
<td>Papanicolaou test for early detection of cervical cancer</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NCI</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>ALTS</td>
<td>ASCUS-LSIL Triage Study</td>
</tr>
<tr>
<td>NLM</td>
<td>National Library of Medicine</td>
</tr>
<tr>
<td>ASCCP</td>
<td>American Society for Colposcopy and Cervical Pathology</td>
</tr>
<tr>
<td>SE</td>
<td>Squamous Epithelium</td>
</tr>
<tr>
<td>CE</td>
<td>Columnar Epithelium</td>
</tr>
<tr>
<td>AW</td>
<td>Acetowhite</td>
</tr>
<tr>
<td>SM</td>
<td>Squamous metaplasia</td>
</tr>
<tr>
<td>SR</td>
<td>Specular Reflections</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MR</td>
<td>Magnetic Resonance</td>
</tr>
<tr>
<td>GAC</td>
<td>Geodesic Active Contours</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>CIE – Lab</td>
<td>Luminance (L) and Chrominance (a,b) of color space CIE – Lab</td>
</tr>
<tr>
<td>GBT</td>
<td>Geometric Bounding Toolbox</td>
</tr>
<tr>
<td>MoG</td>
<td>Mixture of Gaussians</td>
</tr>
<tr>
<td>MDL</td>
<td>Minimum Description Length</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a Posteriori</td>
</tr>
</tbody>
</table>
EM  Expectation Maximization
GEM Generalized EM algorithm
TC  Toggle Contrast
KL  Kullback-Leibler
NM Cut Normalized Mean Cut
WM Cut Weighted Mean Cut
N Cut Normalized Cut
Ag  Agglomerative Clustering
CAg Constrained Agglomerative Clustering
Ag U Agglomerative Clustering with Updates
TP  True Positive
TN  True Negative
FP  False Positives
N,P Negative and Positive samples
IQR Interquartile Range
MSR MultiScale Retinex
US  Unsupervised
SUP Supervised
GT  Ground Truth
Knn K Nearest Neighbor
GKL Gaussian-Kullback-Leibler
STD Standard Deviation
SSM Standard Deviation Scaled by Mean
ESM Entropy Scaled by Mean
ROC Receiver Operating Characteristic
PV  Predictive Value
HSV Hue, Saturation and Value, color feature space
HI  Histogram Intersection
$R$  - Radial distance from the image (cervix) center (Sections 2.1.1, 4.5)
- Rough estimation of the illumination field at every pixel (Section 2.3)
- Segmentation mask of automatically detected region (Chapter 4)

$a, b$  The $a$ and $b$ color channels from the Lab color space

$I$  - Brightness (Section 2.1.2)
- Illumination field (Section 2.3)

$S$  - Color saturation (Section 2.1.2)
- A set of nodes sampled along a curve $C$ (Section 2.2.2)
- Ground truth segmentation mask, marked by an expert (Chapter 4)

$\nabla$  Gradient operator

$I(x, y)$  Image function

$f(x, y)$  A scalar weight function

$\Omega$  Function domain

$\mathbb{R}^+$  A space of positive real numbers

$Re^2$  Two dimensional space of real numbers

$C$  - Curve (Section 2.2)
- Cluster (Section 3.2.3)

$s$  An arc length parameter

$C(s)$  A two dimensional curve with parametrization $s$

$L$  - Curve length (Section 2.2)
- Intensity channel of the CIE $-$ Lab color space

$\Omega_C$  A region inside the curve $C$

$E(C)$  Energy functional of the curve $C$

$E_{data}(C)$  Energy functional based on image data information

$E_{shape}(C)$  Energy functional based on a prior shape information
\( \alpha \) - A parameter that controls the activation sequence of \( E_{\text{data}}(C) \) and \( E_{\text{shape}}(C) \) (Section 2.2)

- A parameter of the \textit{F-measure} (Section 5.3.2)

\( \frac{\delta E(C)}{\delta C} \) The first variation of the functional \( E \) with respect to the curve \( C \)

\( \phi(x, y) \) Level set function

\( \phi_t^{\text{data}} \) Level set evolution based on image data information

\( \phi_t^{\text{shape}} \) Level set evolution based on prior shape information

\( g(x, y) \) An inverse edge indicator function

\( \beta_1 \) Weight of the weighted region term within the active contour framework (term 1)

\( \beta_2 \) Weight of the robust alignment term within the active contour framework (term 2)

\( \beta_3 \) Weight of the GAC term within the active contour framework (term 3)

\( \beta_4 \) Weight of the minimal variance term within the active contour framework (term 4)

\( c_1, c_2 \) Estimated mean intensities within and outside the evolving curve

\( t \) Iteration number

\( k_1 \) Largest principle curvature

\( g_{ij} \) Coefficients of the first fundamental form at pixel position \((i, j)\)

\( b_{ij} \) Coefficients of the second fundamental form at pixel position \((i, j)\)

\( \vec{V} \) Direction vector of \( k_1 \)

\( \tilde{k}_1 \) Normalized \( k_1 \)

\( \tilde{g}(x, y) \) An inverse edge indicator function based on curvature features

\( \Phi \) A continuously defined feature along a curve

\( \Lambda \) The range of values of a feature along a curve

\( \lambda \) A variable spanning \( \Lambda \)
$H(\lambda)$ The cumulative distribution function (CDF) of $\lambda$

$h(x)$ An indicator function

$H(C, \lambda)$ The CDF of a specific curve $C$

$H^*(\lambda)$ The CDF of the prior shape

$\Gamma$ The parameterized curve as a function of the arc-length

$\vec{\Gamma}(s, t)$ A vector with coordinates \{\text{coordinates of } X(t) - X(s), Y(t) - Y(s)\}

$n(s)$ The the outwards normal at coordinates \{X(s), Y(s)\}

$F_{st}$ The normalized inter-node distance between nodes $s$ and $t$

$G_\sigma$ Gaussian filter with variance $\sigma$

$M, N$ Image dimensions

$O(x)$ Order of magnitude of $x$

$H$ Heaviside function

$H_\epsilon$ A smoothed approximation of the Heaviside function

$\delta_\epsilon(\phi)$ Delta function approximation

$w_d(x, y)$ Local weighting function of the shape term

$\gamma$ Weight of the shape term within the active contour framework

$T$ Threshold used in different places

- Texture features (Appendix A)

$K$ Number of interpolation points (Section 2.3)

$\phi_k$ 2-D Lagrange’s interpolation basis function

$c_k$ Illumination field parameter within the set of interpolation points $c$

$x_i$ Vector of the $i$ pixel’s location - ($x_i, y_i$)

$y_i$ Intensity of pixel $i$

$\hat{y}_i$ Reflectance of pixel $i$

$\theta$ Parameter set for a given density function

$\mu_j$ Mean vector of Gaussian $j$

$\Sigma_j$ Covariance matrix of Gaussian $j$
\( \sigma_j \) - Variance of 1d Gaussian \( j \) (Section 2.3)
- Local scaling factor of the self-tuning kernel (Section 3.2.2)

\( \alpha_j \) - Probability of occurrence of Gaussian \( j \)

\( f(y_i|\theta_j) \) - Gaussian distribution of a random variable \( y_i \)

\( f(\tilde{y}_i|\theta, c) \) - MoG density function for a random variable \( y_i \) given a parameter sets

\( T_i \) - Tissue type at position \( i \)

\( J \) - Number of tissues (Gaussians)

\( n \) - Number of pixels

\( A \) - Matrix form of the Lagrange’s interpolation basis function

\( w_{ij} \) - GEM weight assigned to each pixel (Section 2.3)
- MAP-EM weight assigned to each pixel (Section 3.1.2)
- Weight of a graph edge between nodes \( i \) and \( j \) (Section 3.2.2)

\( \tilde{y} \) - Pixel’s predicted intensity value, without illumination field

\( L_c \) - Position of the intensity histogram peak following normalization

\( \theta' \) - Parameter set for a given prior density function

\( \mu_j' \) - Mean vector of prior Gaussian \( j \)

\( \Sigma_j' \) - Covariance matrix of prior Gaussian \( j \)

\( \alpha_j' \) - Probability of occurrence of prior Gaussian \( j \)

\( n_j \) - Number of pixels from tissue \( j \)

\( \beta \) - A scalar parameter of the MAP-EM algorithm

\( \theta' \) - Parameter set for a given prior density function

\( \hat{\mu}_j \) - Mean vector of Gaussian \( j \), estimated by MAP-EM

\( \hat{\Sigma}_j \) - Covariance matrix of Gaussian \( j \), estimated by MAP-EM

\( \hat{\alpha}_j \) - Probability of occurrence of Gaussian \( j \), estimated by MAP-EM

\( \beta_j \) - A scalar parameter that controls the influence of prior Gaussian \( j \)

\( k^2(x, y) \) - Two-stage toggle contrast transform

\( \psi_1(x, y) \) - Morphological erosion operator

\( \psi_2(x, y) \) - Morphological dilation operator

\( G(V, E) \) - A Graph with a set of vertices \( V \) and a set of edges \( E \)
\( w_e(i, j) \) Edge cost (between superpixels)
\( w_r(i, j) \) Region similarity (between superpixels)
\( W \) Similarity matrix between superpixels
\( KL_{ij} \) KL divergence between superpixels \( i \) and \( j \)
\( SKL_{ij} \) Symmetric KL divergence between superpixels \( i \) and \( j \)
\( K \) - Number of nearest neighbors (Section 3.2.2)
- Number of clusters (Section 4.4.3)
\( e_{ij} \) Gradients energy of a superpixel boundary portion
\( \hat{e}_{ij} \) Locally scaled edge cost
\( S(C_1, C_2) \) A graph cut between clusters \( C_1, C_2 \)
\( a_{ij} \) Length (in pixels) of a superpixel boundary portion
\( l_{WMCut} \) Linkage measure of WMCut
\( l_{MinMax} \) Linkage measure of MinMax-cut
\( l_{NMCut} \) Linkage measure of NMCut
\( N \) Number of superpixels parameter of the AgU algorithm (Section 3.2.3)
\( \hat{R} \) Complement mask of the automatically segmented region
\( e_{\text{min}}, e_{\text{max}} \) Minimum and maximum gradient energy within the local neighborhood of a superpixel boundary portion
\( e_{\mu}, e_{\sigma} \) Mean and variance of gradient energy within the local neighborhood of a superpixel boundary portion
\( E^1, \ldots, E^4 \) Local scaling schemes of the edge cost of a boundary portion
\( p, q \) - Discrete histograms of superpixels region features (Section 4.4.2)
- Sensitivity and specificity, respectively (Section 5.2)
\( D_{ij} \) The decision made by expert \( j \) for pixel \( i \)
\( W_i^{(k-1)} \) \( f(T_i | D_i^{(k-1)}, p^{(k-1)}, q^{(k-1)}) \) Unobserved true segmentation via STAPLE
\( T_i \) The hidden value of the true segmentation (tissue) at pixel \( i \)
\( f(T_i) \) Prior probability for tissue \( i \)
\( h_i \) Probability of bin \( i \) of a histogram
\( F \) \( F\text{-measure} \), the harmonic mean
List of Figures

1.1 An example cervigram ........................................... 3
1.2 Cervigrams content variability ................................. 7
1.3 Cervigram analysis framework ................................. 10

2.1 Region-of-interest detection .................................... 14
2.2 SR Identification .................................................. 16
2.3 Curvature-based boundary indicators ......................... 21
2.4 Cervix boundary detection examples: curvature vs. intensity gradient edge indicators ........................................... 23
2.5 Inter-node distances .............................................. 27
2.6 Incorporating prior shape information in the curve evolution functional .............................................. 30
2.7 Local weights effect on boundary detection quality ........ 32
2.8 Intensity histograms for tissues within the cervix .......... 38

3.1 Variability of cervix tissues ................................... 40
3.2 Distribution of tissue samples in $L_b$ feature space ......... 41
3.3 Pixel-based segmentation process .............................. 45
3.4 Superpixels generation process ................................. 49
3.5 Superpixels representation and similarity ..................... 52

4.1 Cervix boundary detection examples .......................... 61
4.2 Cervix boundary detection with different shape priors ........ 63
4.3 A comparison between the circular-prior and the elliptical-prior via box-plots ............................................ 65
4.4 Automated boundary detection, as compared to the markings of two experts ................................................................. 66
4.5 Illumination correction examples .................................................. 69
4.6 Box-plot results for segmentation quality with and without illumination correction ................................................. 70
4.7 Box-plot results of the MAP-EM process for different $\beta$ values ................................................................. 72
4.8 MAP-EM segmentation examples for different $\beta$ values ................................................................. 73
4.9 A summary of segmentation results following illumination correction ................................................................. 74
4.10 Examples for segmentation via clustering of superpixels .................................................................................. 77
4.11 Superpixels accuracy .................................................................................................................................................. 77
4.12 Similarity matrix quality computation .............................................................................................................. 80
4.13 A comparison between different local scaling schemes of the edge cost ................................................................. 81
4.14 A comparison between different similarity matrices via agglomerative clustering .................................................. 83
4.15 Distribution of inter-tissue and intra-tissue similarities .................................................................................. 84
4.16 Agglomerative segmentation example I .............................................................................................................. 86
4.17 Agglomerative segmentation example II .............................................................................................................. 87
4.18 Agglomerative segmentation example III ........................................................................................................... 88
4.19 Average segmentation results for agglomerative clustering with different graph-cut criterions .......................................................... 90
4.20 Box-plot results for agglomerative clustering with different graph-cut criterions .......................................................... 91
4.21 Box-plot results for CAg algorithm with different thresholds .................................................................................. 93
4.22 Box-plot results for final number of segments with CAg algorithm ........................................................................ 93
4.23 CAg segmentation examples ........................................................................................................................................ 94
4.24 AW and CE segmentation results with the AgU algorithm .................................................................................. 96
4.25 A comparison between different clustering algorithms .................................................................................. 97
4.26 Algorithms comparison for a different number of segments .................................................................................. 99
4.27 Segmentation examples with different clustering algorithms .............................................................................. 100
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.28</td>
<td>Classification results with different classification schemes</td>
<td>103</td>
</tr>
<tr>
<td>4.29</td>
<td>Classification results: A comparison between image representations</td>
<td>104</td>
</tr>
<tr>
<td>4.30</td>
<td>Classification results with the addition of context-based rules</td>
<td>105</td>
</tr>
<tr>
<td>4.31</td>
<td>Classification examples</td>
<td>107</td>
</tr>
<tr>
<td>5.1</td>
<td>Examples of manually segmented cervigrams and corresponding multi-</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>expert ground truth</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>STAPLE simulation</td>
<td>114</td>
</tr>
<tr>
<td>5.3</td>
<td>Examples for multi-expert ground truth and corresponding</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>segmentation-complexity descriptors</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Scatter of segmentation complexity in entropy-mean feature space</td>
<td>125</td>
</tr>
<tr>
<td>5.5</td>
<td>Example segmentations for the three groups observed in the entropy-</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>mean feature space</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Distributions of segmentation complexity descriptors in different</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>segmentation tasks</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Examples for segmentation evaluation with different performance</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>measures</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Accuracy results for segmentation in different feature spaces</td>
<td>156</td>
</tr>
<tr>
<td>B.1</td>
<td>Superpixels representation quality via classification - Accuracy</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>results</td>
<td></td>
</tr>
<tr>
<td>B.2</td>
<td>Superpixels representation quality via classification - Miss</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>detection amount</td>
<td></td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Different tissue properties</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Linkage measures for clustering</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Sub-sets within the cervigrams archive</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Quantitative results for boundary detection</td>
<td>60</td>
</tr>
<tr>
<td>4.3</td>
<td>Average results of cervix boundary detection with and without the circular shape prior</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>Average results of cervix boundary detection with circular and elliptical shape priors</td>
<td>64</td>
</tr>
<tr>
<td>4.5</td>
<td>P-values computed for a comparison between the circular-prior and the elliptical-prior</td>
<td>65</td>
</tr>
<tr>
<td>4.6</td>
<td>Average Dice and Sensitivity measures for a comparison between the algorithm and the results of two experts</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>P-values computed for segmentation results following illumination correction</td>
<td>75</td>
</tr>
<tr>
<td>4.8</td>
<td>P-values computed for a comparison between different clustering algorithms</td>
<td>98</td>
</tr>
<tr>
<td>4.9</td>
<td>P-values computed for classification results with the addition of context-based rules</td>
<td>106</td>
</tr>
<tr>
<td>5.1</td>
<td>STAPLE simulation: sensitivity and specificity</td>
<td>115</td>
</tr>
<tr>
<td>5.2</td>
<td>STAPLE simulation: $F$-measure, mean predictive value and accuracy</td>
<td>120</td>
</tr>
</tbody>
</table>
5.3 Evaluation of two algorithms for cervix boundary detection using different performance measures 129
Chapter 1

Introduction

This research is focused on automatic analysis of optical images of the uterine cervix, termed “cervigrams”. Cervigrams are currently being investigated as a means for detection, diagnosis and basic research of cervical cancer. The current work aims at developing and evaluating automated cervigram analysis tools, which are needed for these tasks. We start with an introduction to the clinical perspectives on cervicography and proceed to the image analysis challenges.

1.1 Cervicography in cervical cancer research

Cervical cancer, one of the most common cancer affecting women worldwide and the most common in developing countries (Parkin et al. [72], Eluf-Neto and Nascimento [16]) can be cured in almost all patients, if detected by high quality screening and treated. However, cervical cancer incidence and mortality remain high in resource-poor regions, where high quality Pap (cytology) screening programs often cannot be maintained because of inherent complexity and cost. An alternative method of cervical cancer screening, called visual inspection with acetic acid, is based on color change of cervix tissues when exposed to acetic acid. This inexpensive method helps to detect abnormal cells that turn white (acetowhite) following the application of 3%-5% acetic acid (Wright [44]). An analogous photographic method that permits archive and study is cervicography.
Cervicography was first described by Dr. Staff in 1981 [87]. In this method the uterine cervix is photographed with a special fixed-focus 35mm camera equipped with a ring flash that is used to provide enhanced illumination of the target region. Figure 1.1 shows an example cervicographic image. During the image acquisition process the photographer manually moves the camera back and forth to get the image in focus. The fixed focus of the camera preserves a constant distance between the camera and the cervix, which allows to have comparable pictures of the cervices of all patients and to perform measurements of areas within the cervix. Immediately before the pictures are taken the cervix is washed with 3%-5% acetic acid for one minute. The acetic acid facilitates the removal of any remaining mucus and highlights the abnormal epithelium. Later, an expert in cervical pathology projects those pictures onto a screen to get a magnified image of the cervix, where he can look for different characteristics of the infected or precancerous epithelium. Invasive squamous-cell cervical cancers are preceded by a long phase of preinvasive disease, collectively referred to as cervical intraepithelial neoplasia (CIN). CIN may be categorized into grades 1, 2 and 3 depending upon the proportion of the thickness of the epithelium showing mature and differentiated cells. Visual characteristics such as whitening of the infected regions, the vascular patterns within them and their margins can be used to define the CIN grade (Reid and Scalzi [77]).

For epidemiologic investigations, a cervigram resembles a low-magnification colposcopic image. When additional screening techniques are available, visual methods like cervicography may be used at the initial examination level and patients with indicators of concern are then referred to colposcopic and/or Pap smear screening, or to treatment. Alternatively, DNA testing for a major risk factor for cervical cancer, human papillomavirus (HPV) infection, could be used as a primary screen with visual triage.

The National Cancer Institute (NCI), National Institutes of Health (NIH), has collected a substantial amount of biomedical information related to the occurrence and evolution of uterine cervical cancer in longitudinal multi-year studies carried out in Guanacaste, Costa Rica, and in the United States. The Guanacaste Project is an in-
Figure 1.1: A typical cervigram: marked are the cervix region, the columnar epithelium (CE), the squamous epithelium (SE), the acetowhite (AW), the entrance to the endocervical canal (Os) and the specular reflection artifacts.

tensive 7-year population-based cohort study of HPV infection and cervical neoplasia among 10,000 women in Guanacaste, where the rates of cervical cancer are perennially high. The ASCUS-LSIL Triage Study (ALTS) conducted in the United States, is a randomized clinical trial of management strategies for minor cervical cytological abnormalities, with two years of semi-annual follow-up. State-of-the-art visual, microscopic, and molecular screening tests were used in these studies to examine the origins of cervical pre-cancer/cancer. The Guanacaste and ALTS projects now have a variety of subprojects based on collected specimens, visual images, and outcomes. Data collected includes patient age, sexual/reproductive history, laboratory test results; including Pap smear and cytology, and 100,000 cervicographic images in the form of 35mm color slides, as well as medical classifications for the cervigrams into diagnostic categories (Herrero et al. [32, 33], Schiffman and Castle [81]). NCI along with the National Library of Medicine (NLM) and the American Society for Colposcopy and Cervical Pathology (ASCCP) have formed the NIH-ASCCP Research Group that plans to use these images for the training and education of colposcopic practitioners (Massad [64]). A major long-term objective is to develop a unique Web-accessible database of digitized cervix images for investigating the role of HPV in the development of cervical
cancer and its intraepithelial precursor lesions in women. This work is a collaboration within the NIH between NCI and NLM (Long et al. [59]).

The cervigrams within the NIH archive are unlabeled and have no attached annotation that describes their visual features. Manual extraction of important regions within the cervix by medical experts is not feasible due to the large amount of images within the archive. It is also known to be a subjective process. Important information that relates to the visual content of the cervigrams is therefore currently inaccessible for the different NIH projects and automated analysis of the cervigram images by computerized tools is highly desirable.

The current research targets the generation of an automated analysis framework for cervigrams. Such a framework should provide an accurate delineation of the different regions within the cervix and enable the extraction of various visual features from the segmented regions. Features such as: size, color, texture, shape and relative position within the cervix, can then be used as indices that enable content based access to the NIH archive. These indices can be used for cervical cancer research, to assist in training of experts and to enable future computerized cancer screening. The individual cervigram image and the challenges of the cervigram segmentation task are described next.

1.2 Automated cervigram analysis challenges

A typical cervogram is presented in Figure 1.1. The cervix region, which is the main region of interest within the cervogram, is located in the central part of the image with surrounding vaginal walls and parts of clinical equipment, such as speculum or swab, intruding in the image. A dark surrounding frame can be seen containing lines and text that are overlaid on the image at time of the photographic development. The cervix region is defined by the cervix boundary. Automated detection of the cervix boundary defines the region of medical and anatomical interest within the cervigram and enables further analysis to focus within the cervix region itself. An additional
important landmark is the *os* which is the opening of the cervix. The shape, size and color of the os may vary strongly among cervigrams, but it is usually clearly visible. The os landmark is used by the medical experts as a reference point for interpreting cervix anatomy.

The main tissues of interest that were defined by NCI experts for the automatic cervix analysis task are: 1) The *Squamous Epithelium* (SE) - the normal cervix tissue. It appears as a homogenous pinkish-tan color and consists of multiple layers of cells (Sellors and Sankaranarayanan [82]); 2) The *Columnar Epithelium* (CE), which everts out of the os when the cervix grows rapidly and enlarges under the influence of estrogen (after menarche and during pregnancy). This tissue is very thin (single cell layer) thus the underlying vasculature is in close proximity to the surface. Light transmits easily through the CE tissue and is heavily absorbed in the red spectrum of the blood. Due to the minimal reflectance and maximum absorption the tissue is characterized by a bright red color and a rough textured appearance. The CE region is not always visible within a cervigram image; 3) The *acetowhite* (AW) region is a white-appearing epithelium that is visible for a short period following the application of acetic acid and is a major visual indicator for cervical cancer. Areas of CIN undergo maximal coagulation due to their higher content of nuclear protein and prevent light from passing through the epithelium. As a result, the subepithelial vessel pattern is obliterated and less easy to see and the epithelium appears white. This reaction produces a noticeable effect compared with the normal pinkish color of the surrounding normal squamous epithelium.

Several other regions may appear within the cervix image. Among them are mucus, blood stains and Squamous metaplasia (SM). The SM refers to the physiological replacement of the everted columnar epithelium on the external cervix by a newly formed squamous epithelium. This healthy tissue may appear as white regions that can be confused with AW lesions. The region of the cervix where squamous metaplasia occurs is referred to as the transformation zone. Identifying the transformation zone is of great importance for the visual inspection, as almost all manifestations of cervical carcinogenesis occur in this zone.
Automated analysis of cervigrams is a very complex and challenging task due to a variety of factors:

• Several artifacts are generated during the acquisition process: Due to the strong flash of the camera and convex shape of the cervix the image tends to be brighter around the cervix center and the illumination decreases gradually towards the cervix boundary. This results in inhomogeneous appearance within and across the tissues, which confuses automatic segmentation algorithms. Bright regions may be misclassified as AW lesions, while AW lesions located in the shaded regions won’t be detected. Additional artifacts that interfere with the tissues segmentation are the specular reflections (SR) artifacts. These artifacts are small and bright regions on the cervix surface, which are generated during the image acquisition process due to the presence of fluids (Figure 1.1).

• A large variability is present within the cervigram archive: the image acquisition setup is not constant. The viewing angle varies significantly across the images causing the cervix region to differ in intensity and shape from one image to another. In addition, the physical scene that is imaged has intrinsic variability. For example, in different patients the cervix is not the same size and additional non-cervix tissues or medical instruments may exist. A significant difficulty is the variability of cervix tissue content within the images, as not all defined tissue types are present in each cervigram image. Figure 1.2 shows example cervigrams that illustrate the content variability of the images within the archive.

• Finally, the narrow dynamic range of colors and the lack of distinct boundaries between tissue regions, introduce additional challenging image analysis and data classification tasks.

Initial studies can be found on the analysis of individual cervigram images, or the higher-resolution colposcopic images. Most of these studies are semi-automated, requiring the user to mark regions of interest on various cervix tissues (Cristoforoni et
Features such as color \cite{pogue2016}, texture \cite{ji2015} and shape \cite{ji2015} are then extracted. Based on these features the manually extracted regions are associated with different cervix tissues using various classifiers, such as neural networks \cite{ji2015} or the minimum distance classifier \cite{pogue2016}. Additional works have started to address the task of fully-automated colposcopic image analysis (Lange \cite{lange2017}, Van-Raad \textit{et al.} \cite{vanraad2017}). The data in these works was collected under controlled illumination conditions, thus variations in illumination are minimal. Preliminary segmentation efforts for the cervigrams within the NIH database were recently introduced (Gordon \textit{et al.} \cite{gordon2018}, Srinivasan \textit{et al.} \cite{srinivasan2018}, Xue \textit{et al.} \cite{xue2018}, Huang \textit{et al.} \cite{huang2018}). In these efforts it was concluded that the illumination field inhomogeneity is one of the major causes for difficulties in the analysis \cite{gordon2018, srinivasan2018}. The works to date usually focus on one specific analysis task (e.g. single landmark or tissue) or show initial results with a small number of image examples. Several of the works published in this area, as well as companies working in related fields (Park \textit{et al.} \cite{park2018}, LUMA (TM) Cervical Imaging System \cite{luma}, STI Medical Systems \cite{sti}), are focusing on developing novel acquisition protocols to enforce a more controlled imaging environment and to facilitate the image analysis tasks by utilizing very advanced imaging or fusion of imaging modalities.

\section*{1.3 Main objectives and significance}

The main objective of this research is to develop an automated image analysis framework for the cervigrams within the NIH archive that will enable future content-based indexing and retrieval of these images. The candidate biomedical regions desirable
for indexing include the squamous epithelium, the columnar epithelium and the acetowhite. Detection of additional anatomical features such as the cervix boundary, is of great value. Existing analysis tools, used in previous approaches, are not suited for handling the existing large image repository of the NIH cervigrams. The cervigram acquisition process in a natural, uncontrolled environment, introduces a high degree of complexity into the analysis task that should be considered. The cervigrams contain a high degree of noise, such as specular reflections and a large variability exists within acquired images, in their relative colorings, shadows, and geometrical layout of the various tissues. The developed tools should be robust enough to handle this large variability. The tools should be fully automatic, as even semi-automatic tools are time consuming for such a large database. Finally the tools can rely only on the information available within the cervigrams, which relates to the appearance of the cervix following the application of acetic acid. Additional visual information such as the appearance of the cervix prior to the wash, or acquisition with other imaging modalities, is not available. These are the challenges of the automatic analysis tools developed in this research. It is highly desirable however, that the tools be general enough to facilitate analysis across additional, similar archives or modalities of cervix images.

Specific aims of the current study include:

- Develop processing tools that can accommodate for the noise and illumination variability across the images.

- Automatically allocate and segment the cervix region (boundary) within the cervigram image.

- Automatically divide the cervix region into segments that coincide with the tissues of interest. The feasibility of two segmentation frameworks is being addressed for this task. One is a pixel-based segmentation framework, which utilizes tissue models and the other is an unsupervised framework that is based on merging of small coherent regions on the image plane termed superpixels.
• An additional issue that is being addressed is segmentation evaluation when the markings from multiple experts are available. The objectives in that case are to define measures for the level of agreement between experts, identify the complexity of different segmentation tasks and evaluate automatic segmentation results. A specific focus is placed on the cervigram images.

The current research is one of the first attempts to automatically analyze the cervigrams within the NIH archive. It provides a major step forward in answering the defined tasks and in identifying the overall challenges of a complete cervigram analysis system. A reliable automatic analysis framework is significant in order to:

• Advance automated screening for cervical cancer by using computerized tools that enable early detection of cervical cancer and support procedural treatment worldwide.

• Automated analysis tools are important for training purposes: they provide guidance and help to medical experts and apprentices using cervicography.

• Once visual information is extracted per image, it is then possible to utilize this information to extract images of similar characteristics from within the given NIH archive.

• The extracted visual data can be used in an ongoing research (conducted by NCI) to validate the cervicographic technology for cervical cancer screening. In this research AW lesions detected in cervigrams are correlated with other pathological findings, an existing correlation will prove the method’s validity.

1.4 Dissertation overview

In order to cope with the large variability of the NIH archive, the current research presents an automated, multi-stage framework for the cervigrams analysis. Each stage within this framework targets a specific region within the cervigram, utilizing features
that were found to be effective for the particular processing step and adequate mathematical tools. Parts of the work presented here have been published or submitted for publication [15, 24, 25, 26, 27, 28, 29, 61, 60, 62, 99].

Figure 1.3: Block diagram of an automated multi-stage analysis framework for cervigram segmentation

The overall automated cervigram segmentation system is presented in Figure 1.3. Methods are described in Chapters 2 and 3. Chapter 2 focuses on automated cervical boundary extraction. It includes two additional preprocessing steps of specular reflection elimination and illumination correction. Chapter 3 focuses on the automatic segmentation of tissues within the detected cervix boundary.

- Extracting ROI and removing specular reflection artifacts (Section 2.1):
  A coarse region-of-interest (ROI) is initially extracted from the given input image. The detected ROI is intended to exclude as much irrelevant information as possible, while making sure that the entire cervical area is included (Gordon et al. [28], Greenspan et al. [29]). A second important pre-processing step addresses the problem of specular reflection detection and elimination (Zimmerman and Greenspan [100]).

- Cervix boundary detection (Section 2.2):
  The cervix boundary is extracted more accurately using an active contour framework that incorporates edge, color, and shape information. In this process, the relevant edges in the image are described by the local curvature of the cervical surface (Zimmerman et al. [99], Greenspan et al. [29]). Two methods are presented for incorporating the prior shape information (Lotenberg et al. [61, 60]). Subsequent steps of the process are performed within the extracted cervix region.
• Illumination correction and intensity normalization (Section 2.3):

A per-image illumination correction step follows. This step is based on the generalized expectation maximization algorithm (GEM) introduced by Van Leemput et al. [52] for bias correction in magnetic resonance (MR) brain images. At the end of the illumination correction step the modified intensity histograms are normalized across the images in the database. The normalization is performed using prior knowledge on the intensity distribution of the cervix tissues (Dvir et al. [15], Gordon et al. [24]).

• Two methods are presented for cervix tissues segmentation:

1. Pixel-based segmentation (Section 3.1):

   This method is based on a probabilistic pixel-based segmentation framework that can generate a wide range of tissue models. The models utilize both the data from the input image (the unsupervised information) and the prior tissue models that were learned in advance, in a supervised manner (Gordon et al. [24]). This learning process is performed via a Maximum-a-Posteriori version of the Expectation Maximization algorithm (MAP-EM) for a mixture of Gaussians (Goldberger and Greenspan [22]).

2. Segmentation via clustering of superpixels (Section 3.2):

   A new methodology that enables the segmentation of elongated, thin and non-convex regions within the cervix is presented. The framework transitions from pixels to a set of small coherent regions (superpixels), which are clustered bottom-up into larger, perceptually similar regions. This is done utilizing a new graph-cut criterion (Gordon and Greenspan [25, 26]) and different variations of the agglomerative clustering algorithm. Superpixels similarity in this framework is computed as a combined region and boundary information measure.

Experiments and results for different steps within this framework are presented in Chapter 4. Chapter 5 addresses the task of segmentation evaluation when markings
from multiple medical experts are available (Lotenberg et al. [62], Gordon et al. [27]).
This chapter presents some general concepts, with a focus on the cervigrams archive.
The chapter is self-contained and is added at the end of the book, as it is not di-
rectly related to the automated segmentation framework of the cervigram images. A
discussion concludes the work in Chapter 6.
Chapter 2

Cervigram preprocessing and automated cervical boundary detection

2.1 Extracting ROI and removing specular reflection artifacts

The first two steps of the cervigram analysis framework are pre-processing steps essential to the successful computerized analysis of cervigrams. First, irrelevant image regions are discarded by an automated region-of-interest (ROI) detection algorithm, presented in Section 2.1.1. In the second step, regions with specular reflection (SR) artifacts are detected and eliminated. A special algorithm that was devised for this purpose by Zimmermann and Greenspan [100] is shortly described in Section 2.1.2.

2.1.1 Initial automated ROI detection

The cervix region is a relatively pink region located near the image center. For an initial delineation of the cervix, we use two features: the $a$ color channel of the CIE – Lab color space (the higher the value of $a$, the “redder” the pixel color) and the distance $R$ of a pixel from the image center. The $R$ feature provides spatial information and
supports the extraction of continuous regions within the image plane. The $a$ color channel is initially smoothed in order to eliminate small details and the two features are normalized by their maximum value in order to obtain a similar range of values.

The image is separated next into two clusters in the two-dimensional ($aR$) feature space; we use Gaussian mixture modeling, initialized by a K-means procedure, as a statistical clustering methodology (Bishop [3]). The cluster that has the highest $a$-mean and the lowest $R$-mean is selected. The ROI is chosen as the largest connected component within the pixels associated with this cluster. Post-processing of the ROI includes morphological boundary smoothing and elimination of small holes. Morphological operations are performed with a small structuring element of a fixed size. An example of the clustering process and final ROI is presented in Figure 2.1.

Criteria for judging an acceptable ROI is that it should exclude irrelevant information such as medical equipment, frames, text, and non-cervix tissues (Figure 1.1), while preserving the cervix region in its entirety. These requirements are largely satisfied in all the cervigrams analyzed in the current work. After the ROI has been detected, all subsequent processing takes place in the ROI interior; hence, confusing patterns and textures in the cervigram periphery will be ignored and will not degrade algorithm performance.
2.1.2 Detection and elimination of specular reflection

Specular reflection (SR), or highlights, are small and bright regions on the cervix surface, which are generated during the image acquisition process by strong reflectors, such as fluids on the surface of the cervix (Figure 1.1). These SR artifacts interfere with the content analysis of the regions surrounding them. The bright white regions of SR may be confused with lesions, which are usually relatively bright. Furthermore, the strong gradients created by SR amplify the local contrast, causing potentially erroneous results in any further processing that incorporates gradient or contrast measures. Reliable identification of SR is therefore essential.

Modeling of SR pixels within an image using statistical tools is a difficult task, since the number of pixels belonging to SR regions is extremely small relative to the image size. SR pixels tend to be associated with other clusters, rather than being recognized as a separate cluster. To address this problem, a two-stage process is proposed in order to model the unbalanced clusters. In the first stage, candidate SR regions are detected by using thresholds of high brightness, $I$, and low color saturation, $S$, as suggested by Lehmann and Palm [53]. In the second stage, the candidate SR mask is further refined by selecting only the pixels in the vicinity of strong gradients. Pixels within the candidate regions form a subset of the entire image pixel set, as shown in Figure 2.2(b). Within this reduced set, the amount of SR and non-SR pixels is more balanced, and the two modes can be statistically modeled as a mixture of Gaussians.

Experiments suggest that SR pixels are optimally modeled by two Gaussians. One represents the brightest, almost white pixels, that have lost most of their chromatic information. The second Gaussian represents the less bright pixels that retain some of their original color. The non-SR regions are modeled by two additional Gaussians to account for various tissue types. Thus, an overall mixture of four Gaussians is used to model the candidate regions. Following the modeling stage, pixels corresponding to the two Gaussians with the highest mean intensities are selected as the SR pixels. Figure 2.2(c) shows the final set of SR labeled pixels following the statistical modeling
and segmentation.

![Figure 2.2: SR Identification: (a) Original cervigram (cropped around the ROI); (b) Candidate SR regions (black); (c) Final identification of SR (black).](image)

In the processing steps that follow, the SR pixels are usually ignored, but for some purposes it may be desirable to fill in the “holes”, i.e., to introduce new pixel values for the black pixels in Figure 2.2. In such cases a filling scheme that eliminates the strong gradients associated with the SR, while preserving the original texture is used. In this scheme the average color of the surrounding pixels is propagated into the specular regions, creating a smooth filling of the image. This approach is based on the observation that the highlights (the SR) formed on the moist surface of the cervix are very small. The color underneath the highlights is assumed to be nearly constant and similar to the color of the pixels in the immediate surroundings. An evaluation of the SR detection and elimination process can be found in the work of Zimmermann and Greenspan [100].

### 2.2 Cervix boundary detection

The ROI extracted in the pre-processing step is coarse and often includes large parts of the vaginal region. An additional step is required in order to refine the detection quality of the cervix boundary. Various segmentation methods can be used for this purpose, including region-growing and energy minimization functionals. In region growing (Gonzalez and Woods [23]) a region is defined via propagation of similar neighboring pixels. A region-growing scenario in the cervix boundary detection task can be defined...
as propagating the region from a single point located in the center of the initial ROI.
The selection of adequate features for the propagation is not an easy task, as different images will require a different set of features in order to advance the region from the center of the cervix to its boundaries. The region-growing can be disrupted by other tissues within the cervix. The boundaries generated in this way cannot be restricted by any smoothing or shape constraints.

Segmentation that combines edge and region information can be achieved using an energy minimization techniques via the active contour framework. This framework can be sub-categorized into snakes (Kass et al. [45]) and level set (Osher and Sethian [70]) methods, two different schemes to carry out the contour deformation process. A review and comparison between different energy functionals was recently presented by He et al. [30]. A main conclusion of the review, in which both methods were evaluated on a set of different medical images, was that the integration of forces from different energy functionals may lead to better segmentation results. The main advantages of such methods, as compared to region-growing, are their ability to integrate local and global information and to account for both region and edge features, while preserving smoothness of the boundaries.

In the current work energy minimization via active contours is used in order to refine the initial ROI so that it matches the actual cervix boundaries more closely. The main contribution of the current work is the energy functional used, that consists of forces and features adequate for the task of cervix boundary detection. Region, edge, and prior shape information are all used for this purpose. We use the implicit implementation via level-sets. Implementation via the parametric snake mechanism may be possible as well.

In an active contour framework the image is considered as a function $I : \Omega \to \mathbb{R}^+$ where $\Omega \in \mathbb{R}^2$ is the image domain. The segmentation problem is mathematically formulated as the search for a contour $C : [0, L] \to \mathbb{R}^2$ in the image, which is optimal with respect to some pre-defined integral measure, $E(C)$, also called the energy functional. Formally, this problem is stated as: $C = \arg\min_C E(C)$. In the current work
CHAPTER 2. PREPROCESSING AND CERVIX BOUNDARY DETECTION

The energy functional consists of two terms: a data term and a shape prior term:

\[ E(C) = E_{\text{data}}(C) + \alpha E_{\text{shape}}(C) \]  

(2.1)

The data term, is activated first and evolves the curve according to features derived from the input image. The shape term is added next and better aligns the contour to a predefined model of the cervix shape. The \( \alpha \) parameter is a time dependant parameter that controls the activation sequence of the two terms.

In the following sections, a full description of the curve evolution process is presented. The data term is described in Section 2.2.1. The shape term is described in Section 2.2.2, where two methods for incorporating prior shape information are presented. The combination of the data and shape terms is discussed in Section 2.2.3.

2.2.1 Curve evolution based on image data

The data term is based on the following general integral measure:  

\[ E_{\text{data}}(C) = \int_C g(C(s))ds + \int \int_{\Omega_C} f(x, y)dxdy, \]

which imposes constraints on the contour, as well as on region properties inside and outside the contour.

The extremals of the chosen energy functional are identified by the Euler-Lagrange equation:  

\[ \frac{\delta E(C)}{\delta C} = 0 \]

and are found by gradient descent in a level set implementation (Osher and Sethian [70]). In this level set formulation a closed curve \( C \) is represented implicitly by embedding in a higher dimensional function \( \phi(x, y) \), where \( C \) is its zero set, \( C = \{ x, y : \phi(x, y) = 0 \} \).

The data term in the current work is based on a curve evolution functional that incorporates edge and region-based information, as suggested for fast edge integration by Kimmel [46]. The compact representation of the level set formulation is the following curve evolution equation:

\[ \phi^{\text{data}}_t = [\beta_1 f(x, y) + \beta_2 \text{sign}(\langle \nabla \phi, \nabla I \rangle) \Delta I + \beta_3 \text{div} \left( g(x, y) \frac{\nabla \phi}{|\nabla \phi|} \right) + \beta_4 (c_2 - c_1) \left( I - \frac{c_1 + c_2}{2} \right)] |\nabla \phi|, \]  

(2.2)
where $I$ is a gray level function of the image, $\phi$ is the level set function and $\beta_1, \beta_2, \beta_3, \beta_4$ are tuning parameters.

The first term of Equation (2.2) is the weighted region term, which advances the curve according to a scalar weight function, $f(x, y)$. The second, robust alignment term, influences the solution curve to align with edges within the image. The third term, which we refer to as the GAC term, is derived from the theory of Geodesic Active Contours. It attains low values for the portions of the contour that overlap image edges, thus preventing these portions from further evolution. In this term, the function $g(x, y)$ is an inverse edge indicator, usually of the form: $g(x, y) = \frac{1}{\sqrt{1 + |\nabla I|^2}}$.

The fourth and final term is the minimal variance term, which attempts to separate the foreground and the background of the image with respect to their relative mean values. The two constants, $c_1, c_2$, are calculated as the mean intensities in the interior and the exterior of the contour, respectively.

A direct application of the level set formulation of Equation (2.2) for the detection of cervix boundaries was experimentally found to be ineffective; we attributed this to the presence of irrelevant edges within the cervigram that are formed by folds of surface tissue and by the various tissues within the cervix. These edges interfere with the curve attraction to the cervix boundaries. For this reason, we have modified the original formulation of Equation (2.2) to adapt to the unique characteristics of the cervigrams. We note again here the importance of the initial pre-processing to crop the image to the coarse ROI. This ensures that the strong edges of the image frame and of the medical instruments do not attract the evolving curve. The curve $C$ is initialized as the boundary of the coarse ROI, which is always larger than the desired final contour. The weighted region term (term 1) is defined as: $f = -\frac{1}{t}$, with $t$ being the iteration number, so that it will have decreasing influence over time, i.e., as the number of iterations increases. This force is set negative to ensure that the contour moves inward from its initial state. The minimal variance term (term 4) uses the $a$ color channel (of the CIE – Lab color space), to represent the pink color of the cervix region. It is assigned a very low weight to reflect the fact that the color difference
between the interior and exterior of the cervix is usually not significant.

The multitude of irrelevant edges in the cervigram image makes the gradient-based
terms (robust alignment and GAC, terms 2 and 3) inappropriate for the task of cervix boundary detection. We propose an alternative edge indicator, based on the cervix convexity. Using convexity as a characterizing feature is motivated by the observation that most of the cervix boundaries are outlined by folds of surface tissue that form narrow valleys and are distinctively concave. The boundaries are easily detected by their largest positive principle curvature, \( k_1 \). Edges generated by the color transition between two different tissue types are expected to have strong intensity gradients but low curvature; thus, the presence of different tissue types is not expected to interfere with the curve evolution for cervix boundary detection.

The principle curvatures measure the maximum and minimum bending of the image surface, \( I(x, y) \), at each point (Farber [17]). The two principal curvatures and directions are obtained as the eigenvalues and eigenvectors, respectively, of the following matrix:

\[
\begin{pmatrix}
g_{22} & -g_{12} \\
-g_{12} & g_{11}
\end{pmatrix}
\begin{pmatrix}
b_{11} & b_{12} \\
b_{12} & b_{22}
\end{pmatrix},
\]

(2.3)

where \( g_{ij} \) and \( b_{ij} \) are related to the first and the second fundamental forms, respectively, and are given by:

\[
\{g_{11}; g_{12}; g_{22}\} = \{I_x^2 + 1; I_xI_y; I_y^2 + 1\}
\]

(2.4)

\[
\{b_{11}; b_{12}; b_{22}\} = \frac{1}{\sqrt{1 + I_x^2 + I_y^2}}\{I_{xx}; I_{xy}; I_{yy}\}.
\]

(2.5)

The largest principle curvature, \( k_1 \), and its direction \( \vec{V} \) are next used for the generation of the curvature-based vector field of the new edge indicator. The \( k_1 \) feature describes both concave (inward bending) and convex (outward bending) regions, corresponding to positive and negative values respectively. The cervix boundaries are concave; therefore, we adapt the \( k_1 \) feature to favor detection of only concave regions. Specifically, we normalize \( k_1 \) to lie in the range between 0 and 1, such that concave regions will have \( k_1 \) near 1, and convex regions will have \( k_1 \) near 0. This adapted \( k_1 \)
is defined by:

\[
\tilde{k}_1 = \frac{(k_1 - \min(k_1))}{(\max(k_1) - \min(k_1))}
\] (2.6)

The unit vector of the principle direction \( \vec{V} \) is scaled by the associated \( \tilde{k}_1 \), thus emphasizing directions of concave regions and suppressing directions of convex regions. The curvature measures are computed on a smoothed version of the input cervigram in order to eliminate small-scale variations. The resulting boundary indicators are illustrated in Figure 2.3. In Figure 2.3(b) and (c) the scaled principle directions are overlaid on a map of the normalized principle curvatures. A correlation can be seen between regions with strong \( \tilde{k}_1 \) values (bright) and small \( \tilde{k}_1 \) values (dark), with concave and convex cervigram regions, respectively.

Figure 2.3: Curvature-based boundary indicators. (a) Smoothed intensity image of a cropped cervigram; (b) Scaled principle directions overlaid on a map of the normalized principle curvature; Concave regions - bright; Convex regions - dark; (c) Zoom-in on one of the strong edges in (b).

We propose to use the normalized curvature feature, \( \tilde{k}_1 \), and its associated principle directions, \( \vec{V} \), to modify the gradient-based terms (robust alignment and GAC) of Equation (2.2). The GAC inverse edge indicator function \( g \) is replaced by:

\[
\tilde{g}(x, y) = \frac{1}{\sqrt{1 + \tilde{k}_1^2}}
\] (2.7)

The robust alignment term is replaced so that it is driven by the vector field of the principle curvature direction \( \vec{V} \), scaled by the normalized curvature. The modified
alignment term is:

\[
\text{sign} \left( \langle \nabla \phi, \tilde{k}_1 \tilde{V} \rangle \right) \text{div}(\tilde{k}_1 \tilde{V}) |\nabla \phi| \tag{2.8}
\]

The new curvature-based level set formulation of the data term is:

\[
\phi_{i}^{data} = \left[ -\frac{\beta_1}{4} + \beta_2 \text{sign} \left( \langle \nabla \phi, \tilde{k}_1 \tilde{V} \rangle \right) \text{div}(\tilde{k}_1 \tilde{V}) \right.
\]
\[
\left. + \beta_3 \text{div} \left( \tilde{g}(x, y) \frac{\nabla \phi}{|\nabla \phi|} \right) + \beta_4 (c_2 - c_1) \left( a - \frac{c_1 + c_2}{2} \right) \right] |\nabla \phi|, \tag{2.9}
\]

where \( a \) is the \( a \)-color channel of the CIE – Lab color space.

Figure 2.4 shows the advantage of using curvature features over the more commonly used intensity gradients. Two examples (I and II) are presented. The input images with the ground truth regions marked by experts are shown in (a). The manually marked cervix boundary, that serves as our ground truth, is marked in yellow. The automatic cervix boundary, found with the curvature-based functional, is marked in red on the original image, as shown in (b). The same curvature based contour, marked on the principal curvature function, is presented in (d). The automatic cervix boundary, found with the gradient-based functional, is marked in red on the original image, in (c). The same gradient-based contour, marked on the intensity gradient, is presented in (e). The boundary of the coarse ROI, that serves as the initial condition for curve evolution, is marked in green in (b) and (c).

In example I, the AW lesion (blue) causes a color transition between the tissues. This transition introduces an intensity gradient (e), which attracts the solution curve in (c). The curvature feature (d) does not exhibit positive response to this particular edge, and thus the resulting curve in (b) remains aligned to the cervix boundary which is well defined by a prominent response in the curvature function.

In example II, the cervix boundary is defined by valleys that yield a very strong positive response in the curvature function (d). The curvature-based contour closely follows this response, resulting in an accurate cervix segmentation (b). The gradient based curve, on the other hand, is easily distracted by the regions that have stronger gradient magnitudes then the actual cervix boundary (c).
Figure 2.4: Two cervigram examples (I) and (II). (a) Original cervigrams with tissues of interest marked. (b) Cervix boundary found using the curvature-based edge indicator (red). Initial coarse ROI boundary (green). (c) Cervix boundary found using the intensity gradient edge indicator (red). Initial coarse ROI boundary (green). (d) Curvature-based boundary contour (red) overlayed on the principal curvature function. Edges with greater relevance attain high values (white). (e) Intensity gradient function overlayed with the gradient-based boundary contour (red).
It is important to note that the curve evolution process is governed by all of the terms in the defined energy functional (Equation (2.9)). Two of the terms are closely related to the concavity measure. When no strong concavity response is present in the vicinity of the curve, it is influenced by the minimal variance and the weighted region terms. This results in a boundary that mostly, but not entirely follows the concave valleys.

2.2.2 Curve evolution based on a prior shape model

Using the data term of the energy functional (Equation (2.9)) the detected cervix boundary is optimal with respect to color and curvature-based edge features. When observing human expert markings, an evident commonality is that the marked boundary is smooth and circular or elliptical in shape. The curve evolution framework is next adapted in order to accommodate this prior shape information and further refine the detection quality of the cervix boundary.

A vast amount of work had been done to embed prior shape information into a segmentation task. A popular approach is to use prior models based on allowable deformation of a template shape. In this approach the shape model is generated from an aligned training-set and different tools are used for its representation. Cootes et al. [9] presented the Active Shape Model, which uses the statistics of a large set of points, sampled at meaningful locations along the object boundary. The generated high-dimensional model requires a vast amount of computational power, thus dimensionality reduction is desired. Leventon et al. used Principle Component Analysis (PCA) in order to extract the mean shape and the eigenshapes and represented them with a Gaussian distribution [55]. In a different work of that group, the model is represented by a level-set function, using intensity and curvature as a function of the signed distance from the object boundary [54]. Following model generation there are works that use an active contour framework in a level-set implementation for the segmentation task (Chen et al. [8], Rousson and Paragios [80]). In these works, registration and segmentation are performed simultaneously to align a new image to the prior
CHAPTER 2. PREPROCESSING AND CERVIX BOUNDARY DETECTION

model. Methods, such as those mentioned above, are inappropriate for the cervix region segmentation task due to the large variability of cervix shapes that exists across the database. Such variability leads to a model with a large range of allowable deformations, that has a minor effect on the resulted segmentation. In addition to that, cross-image registration of cervigrams from different patients is a question that wasn’t addressed thus far, thus a model learned from a registered training set is currently not available.

The objective of the current task is to incorporate elliptical or circular shape information into our active contour framework, using a low-complexity scheme that can cope with the large variability of cervix shapes. Two methods are presented:

I. A method that utilizes the distribution of shape features and a circular shape prior.

II. A novel method that uses an implicit shape representation via a level-set function and an elliptical shape prior.

I. Embedding a circular prior using the distribution of shape features

The first method for embedding shape information in the active contour framework is based on the distribution of shape features (Litvin and Karl [58]). The shape energy term in this method penalizes the differences between the feature distributions of a given curve and of a prior reference curve. It can be shown that such distributions capture the intuitive similarity of shapes in a flexible way, while being invariant to shape transformations (Osada et al. [69]). This method is applied here with a circular shape prior and is hereon termed the circular-prior.

We start with a brief description of the method. We then present specific modifications used for the cervix region segmentation task. The shape distribution is defined as the Cumulative Distribution Function (CDF) of feature values measured uniformly along the shape boundary. Let \( \Phi \) be a continuously defined feature along the curve \( C \), and \( \lambda \) be a variable spanning the range of values \( \Lambda \) of the feature. The CDF of \( \Phi \),
$H(\lambda)$, is defined as:

$$H(\lambda) = \frac{\int_C h\{\Phi(C) < \lambda\}dw}{\int_C dw}. \quad (2.10)$$

Here $h(x)$ is an indicator function, which is 1 when the inequality is satisfied and 0 otherwise. When it is meaningful to exhibit the particular curve $C$ for which $H(\lambda)$ is computed, we will write $H(C, \lambda)$.

The shape energy term, $E_{\text{prior}}(C)$, is defined as:

$$E_{\text{prior}}(C) = \int_{\lambda} [H^*(\lambda) - H(C, \lambda)]^2 d\lambda, \quad (2.11)$$

where $H(C, \lambda)$ is the feature distribution function of the curve $C$, and the prior shape information is captured in the target distribution $H^*(\lambda)$.

The shape descriptor used in the current work to describe the circular shape prior is termed the "inter-node distances" descriptor. This descriptor captures the CDF of the normalized distances between all nodes within the set of nodes $S$, sampled uniformly along the curve. The curve evolution minimization equation for the inter-node distances feature, as defined by Litvin and Karl [58], is:

$$\nabla E(\Gamma)(s) = 2 \int_{t \in S} n(s) \cdot \frac{\vec{\Gamma}(s,t)}{|\vec{\Gamma}(s,t)|} [H^*(F_{st}) - H(\Gamma, F_{st})] dt, \quad (2.12)$$

where $\Gamma$ is the parameterized curve as a function of the arc-length $\{X(s), Y(s)\}$ with $s \in \{0, 1\}$, $\vec{\Gamma}(s,t)$ is a vector with coordinates $\{X(t)-X(s), Y(t)-Y(s)\}$ and $n(s)$ is the outwards normal at $\{X(s), Y(s)\}$. The normalized inter-node distance between nodes $(s,t)$, is defined as:

$$F_{st} = \frac{|\vec{\Gamma}(s,t)|}{\text{mean}(\{|\vec{\Gamma}(s,t)| \mid (s,t) \in S\})}. \quad (2.13)$$

Figure 2.5 illustrates the inter-node distances between a single point, $s$, and the rest of the points, $t_1, \ldots, t_7$, within the sampled set $S$. The histogram of the inter-node shape descriptor, $h(F_{st})$, is constructed out of all the available (normalized) distances between such node pairs.

In the current work we use the following level set implementation for the shape-
Figure 2.5: Inter-node distances between points within the sampled set $S$, on the prior circular model.

Based curve evolution:

$$\phi_t^{\text{shape}} = \nabla E(\Gamma)(s) * G_\sigma(x, y),$$

(2.14)

where $\nabla E(\Gamma)(s)$ is computed for pixels along the zero level set (the evolving curve $C$) and is diffused using a simple Gaussian kernel ($G_\sigma$ in Equation (2.14)) of size $5 \times 5$ and $\sigma = 0.5$.

The circular-prior method has the following advantage that serves the current application: There is no need for registration and alignment of the prior shape model and the evolving curve as the method is based on the distance between normalized feature distributions. These distributions are invariant to scale and rotation. On the other hand, the method is based on a non-parametric sampled curve, and its level-set implementation is not straightforward. Considering an image of size $M \times N$ pixels, and assuming a sampled curve of size $O(x)$ where $x < M \times N$, the algorithm complexity is $O(x^2)$. Originally this method was designed to deal with much more complex shapes, and as such when applied to the cervix boundary segmentation with a prior shape of a simple circle, it might seem too complex.

II. Embedding an elliptical prior using an implicit shape representation via a level-set function

A second, novel method that incorporates the prior shape of an ellipse into the active contour framework, is presented next. As the parameters of the ellipse are not known in advance, the method simultaneously estimates the prior model and per-
forms the segmentation. This is done by alternating between the two steps of elliptical model estimation and cervix boundary segmentation. The method is hereon termed the elliptical-prior.

The elliptical model is generated using the bounding and bounded ellipses of the evolving curve. It is computed using the following procedure: 1) The bounding ellipse is computed using the Geometric Bounding Toolbox for Matlab (GBT\(^1\)); 2) The bounded ellipse is computed using the covariance matrix of the coordinates along the evolving curve. An initial bounded ellipse is generated using a 2D projection of the covariance matrix onto the image plane at a fixed sigma. This initial projection is enlarged until an intersection with the curve is reached; 3) A set of shortest paths (lines) between the pixels of the bounding and bounded ellipses, is extracted; 4) The elliptical prior is defined as the centerline crossing the lines within this set.

The shape energy term of the active contour framework within the segmentation step, minimizes the integral of the non-overlapping areas between the elliptical prior and the evolving curve. This is done implicitly, using the level set representations of the two curves, as suggested by Riklin-Raviv et al. [78]:

\[
E_{\text{shape}}(\phi) = \int_{\Omega} (H_\epsilon(\phi(x)) - H_\epsilon(\tilde{\phi}))^2 dx, \tag{2.15}
\]

In this equation \(\tilde{\phi}\) is the level set of the prior model, \(H\) is the Heaviside function of a given level set and \(H_\epsilon\) is its smoothed approximation (Chan and Vese [7]):

\[
H_\epsilon(\phi) = \frac{1}{2} \left( 1 + \frac{2}{\pi} \arctan \left( \frac{\phi}{\epsilon} \right) \right). \tag{2.16}
\]

The Heaviside function is an indicator function for object and background and is used in that case to define the non-overlapping areas between the evolving curve and the model.

The shape term within the curve evolution equation is obtained by minimizing \(E\)

\(^1\)http://sysbrain.com/gbt/
with respect to $\Phi$:

$$
\phi_{t}^{\text{shape}} = \delta_{t}(\phi)(H_{\epsilon}(\phi) - H_{\epsilon}(\tilde{\phi})),
$$

(2.17)

where $\delta_{t}$ is approximated by:

$$
\delta_{t}(\phi) = \frac{dH_{\epsilon}}{d\phi} = \frac{1}{\pi \epsilon^{2} + \phi^2}.
$$

(2.18)

This method is much simpler to implement (as compared to the circular-prior method) due to its level set formulation. For an image of size $M \times N$ pixels and curve of size $O(x)$, the algorithm complexity is $O(x)$. This complexity is attained by using a narrow band implementation of the level-set function. The model estimation procedure described above is simpler as compared to other methods that estimate general-shape priors (Riklin-Raviv et al. [78]). The elliptical shape prior can be computed directly from the curve coordinates, thus the complex search for optimal model transformation parameters is unnecessary.

### 2.2.3 Combination of the data and shape terms in the curve evolution process

The data and shape terms, $\Phi_{t}^{\text{data}}$ and $\Phi_{t}^{\text{shape}}$, are combined in a two-stage procedure: The first stage uses the data term to evolve the initial, approximate ROI to better fit the cervix region. The curve that is output from this stage will be hereon termed the data-driven curve. The second stage combines the data and shape terms to further refine the detected boundary and obtain a smoother, more circular shape.

This two-stage procedure was found necessary by experimentation: if the two terms are used simultaneously to evolve the curve in a single-stage procedure, the resulting cervix boundary is larger than desired. The data term is attracted to local concavities which are present between the initial ROI and the actual cervix boundary. The shape term tries to generate the optimal bounding circle that includes these concavities, thus preventing the curve from following the data inwards in favor of a more circular shape.
A comparison between the two-stage procedure and the simultaneous activation of the data and shape terms, in a single stage, is presented in Figure 2.6. Figure 2.6(a) shows the medical expert-marked boundary (blue). Figure 2.6(b) shows the initial ROI (green), along with the boundary detected using data and shape terms in a single-stage procedure (red) and the boundary detected using the two-stage procedure (white). Local concavities can be found between the initial ROI curve and the desired cervix boundary (bright values in Figure 2.6(c)). The curve generated by the single stage procedure (red) is attracted to the concavities that are closest to the initial ROI, with the shape term enforcing a circular shape based on these concavities. It was empirically found that parameter tuning of within-term forces and between the two terms does not improve this outcome. The suggested two-stage procedure (white curve) results in a curve that matches more closely with the expert markings.

Figure 2.6: Incorporating prior shape information in the curve evolution functional. (a) Manual marking of the expert (blue); (b) Cervix boundary results: initial ROI (green); boundary detected using data and shape terms in a single-stage procedure (red); boundary detected using the two-stage procedure (white); (c) Curvature feature map (bright regions correspond to local concavities).

An important issue is the relative weighting of the data and shape terms in the second stage of the curve evolution process. Experiments have shown that equally weighting the two terms causes the data-driven contour to overinflate as it tries to match the circular prior model. In order to avoid this undesirable effect, the contour is restricted from evolving outwards beyond a pre-defined distance limit from the data-driven curve (which is in good proximity to the desired cervix boundary). This is done
by using the following equation:

\[ \phi_t = \phi_{data}^t + w_d(x, y)\phi_{shape}^t, \]  

(2.19)

where the function, \( w_d(x, y) \), locally weights the shape term per pixel. The local weights are defined using the signed distance transform, \( d(x, y) \), that computes the minimal distance between a pixel \((x, y)\) and the data-driven curve (positive distances interior to the curve and negative distances exterior to the curve). The weighting function is defined as:

\[ w_d(x, y) = \begin{cases} \frac{\gamma}{d(x,y)} & \text{if } d(x, y) < T \\ \gamma & \text{otherwise} \end{cases}, \]  

(2.20)

where \( T \) specifies the distance limit and is assigned a small value of \( T = -3 \) pixels, and \( \gamma \) is a parameter that controls the general influence of the shape term. The proposed weighting function suppresses the influence of the shape term on pixels positioned outside the data-driven curve. Note that the weights are computed once per data-driven curve.

Figure 2.7 illustrates the effect of the local weights on the boundary detection quality. The data-driven curve is marked in green. The result for equally weighting the shape term is marked in red. The result for locally weighting the shape term, is marked in white. The manual markings of the expert are presented in blue. It can be observed that the equally-weighted curve (red) inflates the input data-driven curve (green). The locally-weighted curve (white) is smoother and better resembles the expert markings. This curve is shown to outperform the data-driven curve, thus supporting the inclusion of the shape term in the curve evolution framework.

### 2.3 Illumination correction and intensity normalization

A standard method for illumination correction is to estimate the illumination field as a single model derived from the lighting conditions and the shape of the object.
Figure 2.7: Local weights effect on boundary detection quality for two cervigrams examples (1)-(2). Left image - manual markings of the expert (blue). Right image - boundary detection results: Data-driven curve (green); Equally weighted shape term (red); Locally weighted shape term (white).

(Gonzales and Woods [23]). Lighting conditions can be computed from a training set of images with known intensity values under normal lighting (Xie and Lam [96]). Shading artifacts generated due to the three-dimensional shape of an object can be corrected using Lambert’s Cosine Law when lighting conditions are known (Jain et al. [37], Klette et al. [48]). These methods require a training set of images which is not available for the NIH cervigrams database. Furthermore the large diversity of shapes and viewing angles across the images make these methods impractical and the illumination field needs to be learned for each cervigram separately. Several works that address a per-image illumination correction can be found in the literature, among them is the state-of-the-art Retinex algorithm for image enhancement (Jobson et al. [41], Kimmel et al. [47], Rahman et al. [76]). The Retinex algorithm is known to provide a strong dynamic range compression and color constancy. It is based on the assumption that the illumination field is spatially smooth. No prior knowledge that relates to the image content is required for the illumination field estimation.

The illumination correction method used in the current work is based on the assumption that the cervix content can be modeled by a mixture of Gaussians. A per-image illumination field correction is thus performed utilizing this assumption and iteratively interleaving segmentation with illumination field estimation. The presented algorithm is based on the generalized expectation maximization algorithm (GEM) introduced by Van Leemput et al. for bias correction in magnetic resonance (MR) brain.
Following is a description of the GEM application to cervigrams: Cervigram tissues are described in the CIE−Lab color space. The illumination correction process is applied to the L channel which represents the intensity levels of each pixel and is the only channel influenced from illumination changes (Gonzales and Woods [23]). Channels a and b remain unchanged. The L channel is initially down-sampled. This slightly improves the results while considerably reducing the running time. As the illumination field is assumed to be multiplicative [23], a logarithmic transformation is performed on the L channel in order to make it additive and to simplify the computations.

The illumination changes gradually within the image plane, thus it can be modeled by a polynomial. In the current work we use the polynomial-type Lagrange interpolation (Steffensen [88]), but the presented theory is valid for any kind of smooth basis function (Van Leemput et al. [52]). We use the following interpolation:

\[
I = \sum_{k=1}^{K} c_k \phi_k(x),
\]

where \(c_k\) are the illumination field parameters, \(\phi_k\) are 2d Lagrange’s interpolation basis functions and \(K\) is the number of the interpolation points (parameters). The interpolation points are located on an equally spaced grid of the image size with \(N \times N = K\) points at positions \((x_v, y_\mu)\) of the image spatial coordinates. The Lagrange’s 2d interpolation-formula is:

\[
I = \sum_{v=1}^{N} \sum_{\mu=1}^{N} Q_v(x_v) \frac{\bar{Q}_\mu(y)}{Q_v(x_v) \bar{Q}_\mu(y_\mu)} c(x_v, y_\mu),
\]

where the following definitions are set: \(Q(x) = (x - x_1)(x - x_2)\ldots(x - x_N)\), \(\bar{Q}(y) = (y-y_1)(y-y_2)\ldots(y-y_N)\), \(Q_v(x) = \frac{Q(x)}{x-x_v}\), \(\bar{Q}_\mu(y) = \frac{Q(y)}{y-y_\mu}\). The illumination field intensity in the position of interpolation point \(k\) is: \(c(x_v, y_\mu)\) and the associated interpolation basis function is \(\phi_k = \frac{Q_v(x)}{Q_v(x_v)} \frac{\bar{Q}_\mu(y)}{\bar{Q}_\mu(y_\mu)}\).

Using this model, the illumination field can be removed from each pixel in the image using: \(\hat{y}_i = y_i - \sum_k c_k \phi_k(x_i)\), where \(y_i\) is the pixel’s intensity and \(x_i\) is a vector of the pixel’s location.

Once the illumination field influence is removed from the intensity values, the cervix
tissue can be modeled by a mixture of Gaussians in 1d feature space. The Gaussian distribution of a d-dimensional random variable is given by:

$$f(y|\theta_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}} \exp\left\{-\frac{1}{2}(y - \mu_j)^T \Sigma_j^{-1}(y - \mu_j)\right\},$$  \hspace{1cm} (2.22)$$

where \(\mu_j\) is the mean, \(\Sigma_j\) the covariance and \(\theta_j = \{\mu_j, \Sigma_j\}\). The covariance is replaced by the variance \(\sigma_j^2\), for the 1d intensity case.

The overall mixture of Gaussians (MoG) density for \(\hat{y}_i\) is:

$$f(\hat{y}_i|\theta, c) = \sum_{j=1}^{J} \alpha_j f(\hat{y}_i|\theta_j, c), \hspace{0.5cm} f(\hat{y}_i|\theta_j, c) = f(y_i - \sum_k c_k \phi_k(x_i)|\sigma_j, \mu_j),$$  \hspace{1cm} (2.23)$$

where \(f(\hat{y}_i|\theta_j, c)\) is the probability density that the pixel value, \(\hat{y}_i\), was generated by tissue type \(j\), given an illumination field with \(c=\{c_k\}\) parameters. The probability of occurrence of each Gaussian \(j\) is denoted \(\alpha_j\). For a set of \(n\) feature vectors: \(\hat{y}_1, \ldots, \hat{y}_n\), the maximum likelihood estimate of the different parameters can be found by maximizing \(f(\hat{y}|\theta, c) = \prod_i f(\hat{y}_i|\theta, c)\), using the GEM iterative algorithm [52].

Having the current estimation of the parameter set, each iteration of the GEM algorithm re-estimates the parameter set according to the following steps:

- **Expectation step:**

$$p(T_i = j|\hat{y}_i, \theta, c) = \frac{\alpha_j f(\hat{y}_i|\theta_j, c)}{\sum_{t=1}^{J} \alpha_t f(\hat{y}_i|\theta_t, c)}$$  \hspace{1cm} (2.24)$$

- **Maximization step of the MoG parameters:**

$$\alpha_j = \frac{\sum_i p(T_i = j|\hat{y}_i, \theta, c)}{n}, \hspace{0.5cm} \mu_j = \frac{\sum_i p(T_i = j|\hat{y}_i, \theta, c)(y_i - \sum_k c_k \phi_k(x_i))}{\sum_i p(T_i = j|\hat{y}_i, \theta, c)},$$

$$\sigma_j^2 = \frac{\sum_i p(T_i = j|\hat{y}_i, \theta, c)(y_i - \sum_k c_k \phi_k(x_i) - \mu_j)^2}{\sum_i p(T_i = j|\hat{y}_i, \theta, c)}.$$  \hspace{1cm} (2.25)$$
• Maximization step of the illumination field parameters:

\[
\begin{bmatrix}
  c_1 \\
  c_2 \\
  \vdots \\
  c_K
\end{bmatrix} = (A^T W A)^{-1} A^T W R; \quad A = \begin{bmatrix}
  \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_K(x_1) \\
  \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_K(x_2) \\
  \vdots & \vdots & \ldots & \vdots
\end{bmatrix}
\]

(2.26)

\[W = \text{diag}(w_i); \quad w_i = \sum_j w_{ij}; \quad w_{ij} = \frac{p(T_i = j|\hat{y}_i, \theta, c)}{\sigma_j^2}\]

(2.27)

\[R = \begin{bmatrix}
  y_1 - \hat{y}_1 \\
  y_2 - \hat{y}_2 \\
  \vdots
\end{bmatrix}; \quad \hat{y}_i = \sum_j w_{ij}\mu_j / \sum_j w_{ij}, \quad (2.28)
\]

where \(T_i\) is the tissue type at position \(i\), \(A\) is the interpolation matrix representing the geometry of the illumination field, \(W\) is the diagonal matrix holding the sum over all weights, \(w_{ij}\), assigned to each pixel. The predicted pixel intensity value is \(\hat{y}_i\), and \(R\) is a vector that represents a rough estimation of the illumination-field at every pixel.

When the GEM algorithm is applied to cervigrams there are several issues that need to be addressed:

1. Setting the degree of the polynomial that estimates the illumination field. This degree defines the smoothness of the illumination field and has a strong influence on the quality of the residual signal within the \(L\) channel. If the surface is not smooth enough (due to a large polynomial degree), important information that relates to the different tissues within the cervix might be lost, leading to a poor distinction between them. It was empirically found that the surface of the illumination field is best described as a fifth degree polynomial (\(K = 36\) interpolation points).

2. When the GEM is used for bias correction in MR brain images [52], an a-priori anatomical atlas of the brain is used in order to initialize the MoG parameters. As no such atlas is available for the cervigram images, a standard K-means clustering
algorithm is used to estimate the initial values of $\mu_j$ and $\sigma_j$. The $c_k$ values are initially zeroed.

3. The total number of tissues $J$ was selected to be 4. This number roughly represents the SE, CE and AW tissues, which occupy most of the cervix region, and an additional group for the remaining tissues. The illumination field estimation should be indifferent to this number as pixels that possess significantly different intensity values from any of the Gaussian distributions will have low $w_{ij}$ values (Equation (2.27)) and their influence on the illumination field estimation will be small (Equation (2.26)).

4. Another important extension of the GEM algorithm for cervigrams is the way specularities and non-cervix tissues are handled once detected in the preprocessing step. These tissues interfere with the illumination field estimation. The filling-in procedure for the detected SR regions, presented in Section 2.1.2, may generate artifacts of smooth bright regions. These artifacts may further interfere with the correct estimation of the illumination field. The GEM algorithm is thus modified to support SR and non-cervix pixels by assigning them small, close to zero, positive $w_{ij}$ values (Equation (2.27)). This enables to perform the interpolation on the whole image plane with a negligible contribution of these pixels to the illumination field estimation. Note that this modification is necessary; the pixels can not be simply masked out as the interpolation step can not ignore pixels.

Following the GEM process, the resulting illumination field is up-sampled to its original size using bicubic interpolation and is subtracted from the log of the original $L$ channel. The corrected $L$ channel is then transformed back to the intensity range using the inverse log transform. Although the illumination field is well estimated, the segmentation during GEM is not accurate enough, as it is performed on the log of a down-sampled image using the intensity channel alone. Important information required for correct segmentation, is ignored in this scenario. An additional step is therefore
required in order to segment the tissues within the cervix.

2.3.1 Intensity normalization

A large diversity of intensity ranges exists within the NIH image database. A dynamic range normalization step is essential in order to learn global tissue models in a supervised manner and to compare images across the data-set. The histogram matching approach, which can be used for this purpose, is not appropriate for the cervigram images because of their content diversity. Some of the cervigrams have no AW or CE tissues at all and histogram matching to a specific histogram model that possess these tissues is erroneous. A cervigram specific normalization process is therefore proposed. This process is based on the assumption that the SE tissue should possess a similar dynamic range of intensities across the image set, as it is the original cervix tissue.

The normalization process suggested is based on the following set of observations that relate to the squamous epithelium tissue (SE). These observations were drawn from a set of 110 manually marked cervigrams: 1) The SE tissue is always present in the cervix image; 2) The SE tissue possesses a narrow intensity range following the illumination correction step. This observation is based on the mean-standard-deviation (mean-std) results that were computed over the entire image set for each tissue, presented in Table 2.1. The SE tissue is shown to possess the lowest mean-std value; 3) The SE tissue occupies most of the cervix region as reflected by the mean-tissue-size (mean-size) in Table 2.1 (measured in pixels).

<table>
<thead>
<tr>
<th></th>
<th>SE</th>
<th>CE</th>
<th>AW</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean-std</td>
<td>2.8</td>
<td>5.4</td>
<td>3.5</td>
</tr>
<tr>
<td>mean-size</td>
<td>36494</td>
<td>5946</td>
<td>5880</td>
</tr>
</tbody>
</table>

Table 2.1: Tissue properties averaged over 110 images.

Based on these three observations, we conclude that the peak of the entire cervix region intensity distribution is strongly correlated with the peak of the SE region intensity distribution. Figure 2.8 shows example cervigram histograms. The histogram
of the entire cervix region (solid line) is shown along with intensity histograms of the individual tissues (SE - dotted line; CE - dashed line; AW - dash-dot line.). The peaks of the entire cervix region histogram and the histogram of the SE tissue are marked with an X. The correlation between the two can be clearly observed.

In the image normalization process we align the SE histograms across the image set. The SE distribution peaks are used as anchor points in this alignment process. Each cervix histogram is shifted so its peak, \( \max(h) \), is set to a fixed value \( L_c \), using the following equation:

\[
L(x, y) = L(x, y) - \max(h) + L_c. \tag{2.29}
\]

In this work \( L_c = 60 \).

Figure 2.8: Intensity histograms of the cervix region (solid line) and of the individual tissues: SE - dotted line; CE - dashed line; AW - dash-dot line. Cervix region peak and SE tissue peak are marked with an X.
Chapter 3

Cervix tissues segmentation

The task of cervix tissues segmentation is one of the most challenging tasks within the automated analysis framework. This task focuses on three main tissues defined by NCI experts: the squamous epithelium (SE), the columnar epithelium (CE) and the acetowhite (AW) (Figure 1.1). Figure 3.1 presents example cervigrams (top row) along with their manual expert markings (bottom row). The cervix region is outlined in yellow, the CE tissue is outlined in purple and the AW tissue is outlined in blue. These examples illustrate the complexity of the segmentation task.

The separation between the AW and the SE tissues is extremely difficult. The AW tissue can be identified by its color, which is lighter than the color of the surrounding SE tissue and by its boundary, but a large overlap exists between the color distributions of these two tissues and the boundaries between them are not always clearly visible. Note for example the AW region in Figure 3.1(c), which is hardly visible. The appearance of a healthy cervix, specifically of the CE tissue, varies over time and the women parity and hormonal status. In addition, the AW lesions are of varying shape, size and can be located in different places within the cervix region, as illustrated in Figure 3.1. Thus no shape constraints can be imposed to aid the segmentation process and the addition of position constraints is not trivial.

Figure 3.2 presents the distribution of tissue samples, extracted out of 110 cervigrams, in $Lb$ feature space (out of the $CIE − Lab$ color space). Samples extracted out
of each tissue are marked by a different color. This figure shows the large overlap that exists across the different tissues in color feature space.

![Figure 3.1: Variability of cervix tissues. Top row: original cervigrams; Bottom row: manual expert markings. Cervix region outlined in yellow; CE tissue outlined in purple; AW tissue outlined in blue.](image)

Two frameworks were examined for the tissue segmentation task. The first framework is based on a probabilistic pixel-based segmentation algorithm and relies on tissues distribution in a selected feature space (Section 3.1). The second framework presents a new methodology that transitions from pixels to a set of small coherent regions (superpixels), which are grouped into larger, perceptually similar regions. Superpixels similarity in this framework is computed via a combined region and boundary information measure (Section 3.2). The two frameworks are fully described next.

### 3.1 Pixel-based segmentation of the cervix tissues

In this section we address the task of pixel-based segmentation of cervix tissues, utilizing Gaussian models for the tissues representation. We use a broad range of models: shifting from models that are learned in an unsupervised manner to models that are learned in a supervised manner. This shift becomes feasible following the preprocessing steps of illumination correction and intensity normalization. We start with a general
Figure 3.2: Distribution of tissue samples in $Lb$ feature space. Each tissue is marked by a different color. Two dimensional projections of Gaussian models learned per tissue are imposed on the plot (outlined in pink). One dimensional histograms are projected on the $L$ and the $b$ axis.

description of the modeling scheme and the probabilistic segmentation process (Section 3.1.1), describe different model learning and segmentation schemes that are used in this framework (Section 3.1.2) and provide a set of context-based rules, which are used to refine the final segmentation output (Section 3.1.3).

3.1.1 Cervix tissue segmentation using a Gaussian Mixture Model

Probabilistic image segmentation consists of three main steps: 1) a transition from image pixels to a selected representative feature space; 2) learning a model for the desired categories (cervix tissues) and 3) pixel classification based on the generated models. The feature space used in this work is the two dimensional $Lb$ feature space ($Lb$ color channels from the CIE – $Lab$ color feature space). The $a$ color channel was experimentally found unnecessary for the cervix tissue representation and is therefore discarded (a full description of the feature selection process is presented in Appendix A). Each tissue within the cervix is next modeled by a Gaussian in $Lb$ feature space and the entire image is modeled as a mixture of $J = 3$ Gaussians. Specularities, detected
during preprocessing, are masked out as they provide misleading tissue information.

The density of a Mixture of Gaussians (MoG) with \( J \) Gaussians is given by:
\[
f(y|\theta) = \sum_{j=1}^{J} \alpha_j f(y|\theta_j),
\]
where \( \alpha_j \) is the probability of occurrence of each Gaussian and \( f(y|\theta_j) \) is the Gaussian distribution, defined by Equation (2.22). Utilizing the MoG for probabilistic segmentation of the cervix tissues, the labeling of a pixel related to a feature vector \( y \) is chosen as:
\[
label(y) = \arg \max_j \alpha_j f(y|\theta_j) \quad j = 1, 2, 3,
\]
where each pixel is affiliated with the label of the most probable Gaussian.

### 3.1.2 Learning the tissue models

There are several learning schemes by which the tissues MoG can be generated and used for the cervix segmentation task. The first scheme is referred to as unsupervised modeling. In this scheme a new MoG is learned in an unsupervised manner, per input image, and is used for its segmentation. Image pixels are considered to be \( n \) independent data points, \( y_1, \ldots, y_n \), and the MoG parameters, \( \theta \), are learned using the EM algorithm in a maximum likelihood framework (Dempster et al. [12]). At the end of the EM process pixels are classified via Equation (3.1). The pixels associated with the Gaussian with the brightest intensity are detected as candidate AW regions; the pixels associated with the Gaussian with the lowest intensity are detected as candidate CE regions. The remaining pixels are associated with the mid intensity Gaussian and are detected as candidate SE regions.

There are two main weaknesses to the unsupervised scheme in the case of cervixograms. The first weakness is the constant number of Gaussians, \( J \), during the model learning phase, which leads to the association of a Gaussian with the AW or CE tissues even if they don’t exists within the image. This weakness results in false detection of the missing tissues. A common solution to this problem is to select among different (larger) \( J \) values using a standard model selection technique (e.g. via the Minimum
Description Length (MDL) (Duda et al. [14]), but such methods usually don’t hold if the number of Gaussians is small and the features are similar (as in the case of the AW and SE tissues). An additional tissue detection step is therefore required at the end of the segmentation process, in order to handle cases where the number of actual tissues within the image is different from \( J = 3 \). Note that we select \( J = 3 \) in the current work in order to be consistent with the other modeling schemes (presented next) and enable a fair comparison between them. The second difficulty is the tendency of the unsupervised modeling scheme to converge into a local optimum and miss-detect tissues that occupy a small portion of the image (such as small AW regions).

The second scheme is referred to as \textit{supervised modeling}. In this scheme a Gaussian tissue model is learned once using a manually marked training set of samples for each of the cervix tissues. The training set is randomly drawn out of the cervigram database. The prior tissues models are next used for probabilistic segmentation of the cervix tissues within a new image using Equation (3.1). The probability for the tissue’s occurrence, \( \alpha_j \), is assumed to be uniform in the cervigram case. No other assumption can be made with regard to this probability due to the large variability that exists within the database: there are images where the AW or CE regions cover almost the entire cervix and there are images where these regions don’t exist. A major advantage of this method is its ability to cope with cases in which one of the tissues is not present in the image.

The third scheme is referred to as the \textit{MAP-EM modeling}. This scheme enables the combination of the unsupervised image information and the supervised models via a MAP-EM framework (Goldberger and Greenspan [22]). In this framework, an EM-based algorithm estimates the maximum a posteriori (MAP) parameters of a mixture of Gaussians density for a given input image. The algorithm integrates between the prior model, which provides more global information on the tissues properties and the observed, local, image data. This process improves the convergence abilities of the algorithm. The prior tissues model, \( f(y|\theta') \), is learned in advance from a training set, as in the supervised modeling scheme. The MAP-EM framework requires an additional
CHAPTER 3. CERVIX TISSUES SEGMENTATION

scalar parameter $\beta$. This parameter controls the relative influence of the prior model and the observed image data on the new model. When $\beta = 0$ the model is learned only from the observed data, as in the unsupervised modeling scheme. When $\beta \to \infty$, the learned model is the prior tissues model itself, as in the supervised modeling scheme.

A single iteration of the MAP-EM framework (fully derived in [22]) consists of the following two steps:

**Expectation step** (same as the maximum likelihood version of the EM):

$$w_{ij} = p(T_i = j | y_i, \theta) = \frac{\alpha_j f(y_i | \theta_j)}{\sum_{t=1}^{J} \alpha_t f(y_i | \theta_t)}, \quad (3.2)$$

where $T_i$ is the tissue type at position $i$.

**Maximization step**:

$$\hat{\alpha}_j = \frac{n_j + \beta_j}{n + \beta}, \quad \hat{\mu}_j = \frac{\sum_{i=1}^{n} w_{ij} y_i + \beta_j \mu'_j}{n_j + \beta_j},$$

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{n} w_{ij} (y_i - \hat{\mu}_j)^2 + \beta_j ((\hat{\mu}_j - \mu'_j)^2 + \Sigma'_j)}{n_j + \beta_j}, \quad (3.3)$$

where $\beta_j = \beta \alpha'_j$ and $n_j = \sum_{i=1}^{n} w_{ij}$. The scalar parameter $\beta$ can be regarded as the number of virtual samples being extracted from the given prior model $f(y | \theta')$ and added to our observed sample set. The larger this number is, the stronger is the influence of the prior model on the learned model.

### 3.1.3 Incorporation of context-based rules

The results of the pixel-based probabilistic segmentation are often noisy and fragmented. An additional step is required in order to smooth out the segmentation map and improve the tissues detection quality. It includes the following context-based rules that were set using prior knowledge regarding the cervix tissues layout:

1. The area of the AW and CE tissues is larger than a predefined size criterion.

2. The CE tissue can be identified as a set of connected components located around
the center of the cervix region.

The segmentation process in this step is described next and is illustrated in Figure 3.3. An input image is shown in Figure 3.3(a). A pixel-based segmentation result is presented in Figure 3.3(b). In this result AW regions are colored dark-red, CE regions are colored light-blue and SE regions are colored yellow. Regions located outside the cervix region and SR pixels are colored dark-blue.

The first rule is implemented using morphological operators that eliminate insignificant regions by a size criterion and provide regions that are continuous (connected) (Figure 3.3(c)). The size criterion was set to 15 pixels for the AW tissue and to 30 pixels for the CE tissue, as their manual markings were observed to be much larger. The eliminated regions were labeled as SE. The second rule is implemented as follows: The distances between the pixels within the CE mask and the cervix boundary are computed by means of the morphological distance transform. A K-means algorithm is used to cluster the pixels into two groups, one close to the cervix center and one close to its boundary (Figure 3.3(d)). Finally, segments that have a majority of pixels associated with the center cluster are identified as CE. The other segments are classified as SE. This provides the final segmentation map (Figure 3.3(e)).

Figure 3.3: Segmentation process: (a) Original cervigram with expert markings: AW - blue, CE - purple, SE - remaining regions within the yellow line. (b) Initial labeling of tissues following pixel-based segmentation, AW - dark red, CE - light blue, SE - yellow; (c) Elimination of small, insignificant regions within the AW and CE maps; (d) Clustering of CE regions into two groups (green regions are closer to the boundary); (e) Final segmentation map.
3.2 Cervix tissues segmentation via clustering of superpixels

In the previous section we have used a pixel-based segmentation framework that relies on color features. This framework was empirically found to provide very poor segmentation results (Section 4.3). Due to the large overlap of tissues color distributions, only a small part of the AW tissue is accurately detected. The pixel-based segmentation without the additional context-based step is very noisy and each segment is broken into many disjoint components (e.g. Figure 3.3(b)). The framework presented in the current section tries to overcome these weaknesses by adding additional information into the segmentation process. This information relates to the local continuity of pixel features within the image plane. The framework shifts from a pixel-based clustering scheme to clustering of small coherent regions in the image plane termed superpixels. Such a region-based clustering scheme has the following benefits: 1) The superpixels representation augments the pixel-based color feature space to include local region and edge distributions. These distributions are statistically robust descriptors for the local image content. Using them for segmentation provides smooth segmentation maps; 2) Similar superpixels exhibit similar properties and can be grouped into larger, perceptually similar regions within the cervix. The complexity and computational cost of clustering is reduced from the number of pixels to the number of superpixels in this case.

Different region-based clustering schemes exist for segmentation of general images. Hermes et al. [31] for example, use local MoG distributions estimated from local color histograms, for the representation of rectangular regions within an image grid. The MoGs are clustered in a deterministic annealing framework, which provides segmentation results in various scales. In a bottom-up aggregation framework used by Sharon et al. [83], segment fragments of increasing size are detected and merged using their coarse scale properties. In a more recent work by O’Callaghan and Bull [68], the morphological Watershed transform is used to generate the primitive regions. The regions are
then represented by histograms and clustered using a spectral clustering technique that optimizes the weighted-mean cut criterion. Another segmentation method based on spectral-clustering and the normalized-cut criterion is used by Fowlkes and Malik [20]. A multi-level region-based framework for AW tissue segmentation within colposcopy images of the uterine cervix was recently introduced by Lange [51]. This work required a substantial amount of human intervention for parameter tuning and presented initial segmentation results.

Our goal is to develop an unsupervised region-based segmentation framework for cervigrams. No work has addressed this task before in a way that can cope with the large within- and across- image variability of the cervigrams, with a minimal number of parameters and human intervention. The suggested framework utilizes the morphological Watershed transform in order to over-segmented the image into a large set of superpixels (Section 3.2.1). A similarity matrix between superpixels that combines region and edge information is constructed next (Section 3.2.2). The matrix is then used in an agglomerative clustering framework, which utilizes different graph-cut criterions (Section 3.2.3).

Specific objectives that relate to the different steps of this framework are: 1) Provide a map of superpixels that accurately decomposes the cervix tissues into small regions; 2) Compute a similarity matrix that can cope with the large variability of color and edge information that exists within and across the cervix tissues; 3) Explore the performance of an agglomerative clustering framework using different graph cut criterions, in the complicated cervigram scenario. A specific focus is placed on a new graph-cut criterion, termed the normalized-mean cut that was devised for the cervigram images. This criterion enables the generation of segments that capture the elongated and non-convex nature of the tissues within the cervix region. The segmentation in this framework is performed within the cervix region (automatically detected, or manually extracted by the expert) following the preprocessing stages of SR detection and illumination correction.
3.2.1 Superpixels generation

Cervigrams are over segmented into superpixels using the Watershed transform (Soille [85]). The Watershed transform is a morphological segmentation tool that is applied to gray-scale images in order to solve a variety of image segmentation problems. The input image to the transform is regarded as a topographic surface, which is being flooded from its regional minima while preventing the merging of the waters coming from different sources. At the end of the process the surface is divided into a set of catchment basins (image regions) separated by crest lines, or watershed lines (which may coincide with region boundaries). The watershed transform is often applied to the gradient image of the original input (Gonzalez and Woods [23]), in which case it has the tendency to over-segment the image into many small regions. The generated segments in that case possess coherent region features and their boundaries are aligned with the image gradients. These properties are desirable for the superpixels in the current framework.

The superpixels generation process includes the following steps (illustrated in Figure 3.4):

1. SR regions are initially filled-in (Section 2.1.2). The image is preprocessed next to smooth out noise and enhance tissue boundaries. This is done using the morphological toggle-contrast (TC) transform (Soille [85]), which is performed on each of the color channels separately. The two-stage toggle contrast transform is defined as:

$$k^2(x, y) = \begin{cases} 
\psi_2(x, y) & \text{if } \psi_2(x, y) - I(x, y) < I(x, y) - \psi_1(x, y) \\
\psi_1(x, y) & \text{otherwise,}
\end{cases} \quad (3.4)$$

where $\psi_1(x, y), \psi_2(x, y)$ are the morphological erosion and dilation operators, respectively and $I(x, y)$ is the pixel value. The process is performed with a small structuring element (SE): a disk of radius $r = 1$. The toggle-contrast sharpens the image edges without boosting the contrast of structures smaller than the
considered SE. An output image for this step is presented in Figure 3.4(b). Regions outside the automatically detected cervix boundary are masked out (colored black) in this example.

2. A color gradient image is generated next (Gonzalez and Woods [23]). The magnitude of the initial gradients is linearly normalized between 0 and 1.

3. A smoothed version of the normalized gradients and a single threshold generates the regional minima for the Watershed transform. The threshold is set lower than the energy of the image boundaries that need to be preserved (defined experimentally). In the current work it is set to $T = 0.01$.

4. The detected regional minima are post-processed to eliminate regions smaller than 3 pixels and are imposed on the normalized gradient image (Figure 3.4(c)).

5. The Watershed transform is applied to this modified gradient image and over-segments the image into superpixels. Figure 3.4(d) presents the boundaries of the generated superpixels, imposed in black. Visual inspection of these boundaries indicates that the superpixels do not cross any important tissues. In particular, an overlap with the expert marked AW boundaries can be detected.

Figure 3.4: Superpixels generation process: (a) Original image with manual markings of the expert imposed. Cervix region outlined in yellow, AW region outlined in blue; (b) Enhanced original image, regions outside the automatically detected cervix boundary are masked out (colored black), SR pixels are filled in; (c) Color gradients with local minima imposed; (d) Superpixels boundaries imposed on original image (in black).
3.2.2 Region & edge similarity matrix

In a typical graph-based image representation, $G(V, E)$, each vertex is a point in the feature space used, associated with a single image pixel. Each edge, $(i, j) \in E$, is weighted by the pairwise similarity, $w_{ij}$, between nodes $i$ and $j$ (Shi and Malik [84]). The weights define the symmetric $n \times n$ similarity matrix $W$, where $n$ is the number of vertices (image pixels). In the current work the graph representation is shifted from a pixel based representation to a region based representation. Each vertex within the graph is a superpixel within the over-segmented image. Each edge is a boundary portion between neighboring superpixels. Edges are weighted by the similarity of the corresponding superpixels. The size of the similarity matrix is considerably reduced, which is crucial in terms of the computational load of graph-based clustering algorithms. The similarity between superpixels is composed from a region similarity, $w_r(i, j)$, and from an edge cost, $w_e(i, j)$. The overall similarity between two superpixels is defined by the square root of their multiplication: $w_{ij} = \sqrt{w_r(i, j)w_e(i, j)}$. Figure 3.5 is an example of the superpixels representation. The original image is presented in (a). Ground truth boundaries between the different tissues are imposed in black, superpixels boundaries are imposed in green. This example is discussed next.

I. Region similarity

For region similarity computation, each superpixel is modeled by a Gaussian distribution in $Lb$ feature space (a full description of the superpixel representation selection process is presented in Appendix B). Figure 3.5(b) illustrates these Gaussians as colored ellipses within corresponding superpixels. Ellipses of superpixels from the same tissue, possess the same color in this illustration. Region similarity between two neighboring superpixels, $w_r(i, j)$, is measured using the symmetric version of the Kullback-Leibler (KL) (Kullback [49]) divergence between two Gaussians, $SKL_{ij}$, locally scaled
CHAPTER 3. CERVIX TISSUES SEGMENTATION

by a self-tuning kernel (Zelnik-Manor and Perona [98]):

\[ w_r(i, j) = \exp \left( -\frac{SKL_{ij}}{\sqrt{\sigma_i \sigma_j}} \right). \] (3.5)

The \( KL_{ij} \) divergence between two \( d \)-dimensional Gaussian components, \( N(\mu_i, \Sigma_i) \) and \( N(\mu_j, \Sigma_j) \), is analytically computed as:

\[ KL_{ij} = \frac{1}{2} \left( \log \left| \frac{\Sigma_j}{\Sigma_i} \right| + \text{Tr}(\Sigma_j^{-1} \Sigma_i) + (\mu_i - \mu_j)^T \Sigma_j^{-1} (\mu_i - \mu_j) - d \right) \] (3.6)

and the symmetric version is computed as: \( SKL_{ij} = \frac{KL_{ij} + KL_{ji}}{2} \). The scalars \( \sigma_i, \sigma_j \) in Equation (3.5) are the local scaling factors. The \( \sigma_i \) factor is computed as the median \( SKL \) value between superpixel \( i \) and its \( K \) nearest neighbors (\( K \leq 5 \), depending on the actual number of neighboring superpixels). Using this kernel, similarity values of neighboring superpixels located within the same tissue are increased and values of neighbors located across tissue boundaries are decreased. Region similarities are normalized between 0 and 1, where higher values indicate stronger similarity. Figure 3.5(d) presents the superpixels boundaries, color-coded with corresponding region similarity values. The \( w_r(i, j) \) values are inverted in this example for display purpose (higher values indicate weaker similarity). Boundary portions that separate between the different tissues in this example (inter-tissue boundaries) obtain higher values (red colors), as expected.

II. Edge cost

The edge cost between two neighboring superpixels, \( w_e(i, j) \), is defined using the average value (energy) of the normalized gradients along their common boundary: \( e_{ij} \).

Figure 3.5(c) presents the superpixels boundaries, color-coded with the respective \( e_{ij} \) values. An edge separating between two tissues is composed of many such boundary portions with varying energy values (as can be seen in this example). The values of the inter-tissue boundary portions are higher as compared to the values of within-tissue boundary portions (the intra-tissue boundaries). The absolute differences between
these two boundary types are small.

Figure 3.5: (a) Original image with tissues boundaries imposed in black, superpixel boundaries imposed in green; (b) Superpixels representation: region features are represented by Gaussians (ellipses) in color feature space. Ellipses of different tissues possess different colors. (c) Superpixels boundaries color-coded with original gradients energy, $e_{ij}$. Stronger boundaries obtain higher weights; (d) Superpixels boundaries color-coded with corresponding region similarity values, $w_r(i, j)$; (e) Superpixels boundaries color-coded with locally scaled energy values, $\hat{e}_{ij}$; (f) Superpixels boundaries color-coded with the overall similarity values, $w(i, j)$. In figures (d) and (f) values are inverted for display purpose and stronger values correspond to weaker similarities.

The edge cost is therefore locally scaled to emphasize values of the inter-tissue boundary portions and to smooth out the values of intra-tissue boundary portions. The scaled energy, $\hat{e}_{ij}$, is computed as:

$$\hat{e}_{ij} = \begin{cases} 
  e_{ij} & \text{if } (e_{\text{max}} - e_{\text{min}}) < T \\
  \sqrt{(1 - \exp(-\frac{|e_{ij} - e_{\text{min}}|}{e_{\text{max}} - e_{\text{min}} + \epsilon}))} e_{ij} & \text{otherwise},
\end{cases}$$

(3.7)

where $e_{\text{max}} = \max_{(k,l) \in N}(e_{kl})$, $e_{\text{min}} = \min_{(k,l) \in N}(e_{kl})$ and $(k,l) \in N$ are the neighboring boundary portions (connected on both ends of boundary $(i,j)$, including boundary $(i,j)$). The thresholding condition in this equation is used to limit unnecessary scaling in neighborhoods with similar values. Such neighborhoods are located within the cervix.
tissues, far from their boundaries. The threshold in this work is set to \( T = 0.01 \). Note that the presented enhancement is performed on the superpixel level rather than the pixel level, thus encompassing more global information.

Figure 3.5(e) presents the superpixels boundaries, color-coded with the \( \hat{e}_{ij} \) values. The dynamic range of the values is increased as compared to the original energies in (c). The differences between inter and intra-tissue boundaries are increased. Some of the intra-tissue values are zeroed. The final edge cost, \( w_e(i, j) \), is computed by applying the negative transform to the \( \hat{e}_{ij} \) values (to be consistent with the region similarity behavior). Figure 3.5(f) presents the inverted values (for display) of the final similarity measure, \( w_{ij} \). The combination of region and edge information (Figures (d) and (e)) increases the strength of inter-tissue boundaries and smoothes out intra-tissue boundaries, as compared to their independent application. Note for example the boundary portions located along the ground truth tissue boundaries of (a). Their final weights in (f), are shown to increase as compared to their edge-cost values in (e). The final weights of the boundary portions within the AW region in (f) are shown to decrease as compared to their region similarity values in (d). Overall the weights of boundary portions within the AW region in (f) are weaker than the weights on its ground truth boundary.

3.2.3 Agglomerative clustering of superpixels

Given a graph model represented by the pairwise similarity matrix, \( W = (w_{ij}) \), perceptually similar superpixels can be clustered using a graph-cut clustering framework (Shi and Malik [84]) and a variety of clustering algorithms. The current work focuses on a new graph-cut criterion termed the normalized-mean cut (NMCut) and on an agglomerative clustering framework, where clusters are built bottom up to optimize the cut criterion.
I. Graph-cut criterions

The classical cut between two clusters $C_1, C_2$, is defined as:
$$S(C_1, C_2) = \sum_{i \in C_1} \sum_{j \in C_2} w_{ij}$$
and the similarity within a cluster $C_1$ is defined as $S(C_1, C_1)$ [84]. Additional cut criterions can be found in the literature (Ding and He [13]). In an agglomerative clustering framework the selected cut criterion is replaced by a corresponding linkage measure, $\ell$. During each iteration of the agglomerative process, the two clusters $C_p$ and $C_q$, that have the largest pairwise linkage between them are merged.

Table 3.1: Different linkage measures between two clusters.

The $\text{MinMax-cut}$ criterion and the corresponding linkage measure $\ell_{\text{MinMax}}$ (Table 3.1-(1)), are shown in the literature to provide the best clustering results [13]. This criterion is defined as minimizing the classical cut, while maximizing the within-clusters similarities. It is known for its tendency to generate balanced clusters (with similar size). The $\text{weighted-mean cut}$ (WMCut) criterion, minimizes the weighted-mean value of the similarities along the cut. It was recently introduced for clustering of superpixels in general images (O’callaghan and Bull [68]). In the $\ell_{\text{WMCut}}$ linkage measure (Table 3.1-(2)) the weights, $a_{ij}$, are defined as the length (in pixels) of the corresponding boundary portions. The WMCut doesn’t favor clusters with a small number of boundary portions [68], but in agglomerative clustering large clusters tend to form even larger clusters, while smaller clusters tend to be left alone [13].

In the current work a new cut criterion, termed the $\text{normalized-mean cut}$ (NMCut), combines the benefits of the MinMax-cut and the WMCut. This criterion (and the corresponding $\ell_{\text{NMCut}}$ linkage measure, Table 3.1-(3)) minimizes the WMCut while maximizing the within-clusters similarities. Using the NMCut more balanced clusters
are generated, but when several neighboring clusters have similar within-cluster similarities (denominator), the clusters are merged according to the actual boundary energy and not the number of boundary portions being considered (numerator). This observation is of major importance in the case of cervigrams, where neighboring tissues may possess very similar within-cluster similarities and where tissues are elongated and non convex, thus a small number of boundary portions may exist between their sub-parts.

II. An agglomerative clustering framework

In the agglomerative clustering framework presented in this work, each cluster is initialized by a single superpixel. The clusters are built bottom up using different merging schemes. In the simple agglomerative clustering algorithm (Ag) the two clusters $C_p$ and $C_q$, that have the largest pairwise linkage between them are grouped in each iteration. The corresponding superpixels within the segmentation map are merged into larger segments. The pairwise linkage measure between the new cluster and the remaining clusters is updated. This procedure has two main drawbacks: 1) At the end of the process a full hierarchy of segmentations is obtained and no rule can predict which is the hierarchy level with the best segmentation accuracy; 2) As the number of clusters decreases, the Ag algorithm may produce erroneous merging that are not correlated with the image content. These merging are derived from the number of superpixels within each cluster, implicitly used by the linkage measure and not by the actual similarity of the two clusters. Examples for erroneous merging are presented in the results section.

We next present two additional clustering schemes with an agglomerative nature that try to overcome these difficulties.

The first scheme is termed the constrained agglomerative clustering (CAG). The CAG algorithm prevents undesired merging by placing additional constraints on the agglomerative procedure. The algorithm finds the two most similar clusters in each iteration via the linkage measure and merges them only if a predefined “stopping rule” is not fulfilled. The stopping rule, or undesired merging, is defined in the current work as the merging of two clusters whose common boundary portions contain at least one
boundary with a weight, \( w_{ij} \), smaller than a predefined threshold, \( T \). Note that \( w_{ij} \) are the initial weights between neighboring superpixels, these weights are not updated along the clustering process (as opposed to the linkage measure). When two clusters are not merged by the stopping rule, their pairwise linkage measure is set to zero to prevent their future merging. When two clusters are merged, the corresponding superpixels are merged into larger segments and the pairwise linkage measure between the new cluster and the remaining clusters is updated (similar to the Ag algorithm). At the end of the CAg process the remaining clusters can no longer be merged. The number of clusters is therefore an output of the algorithm. This number is governed by the stopping rule used and the threshold \( T \).

The second scheme is termed *agglomerative clustering with updates* (AgU). In this scheme whenever one of the clusters reaches a predefined amount of components (superpixels), \( N \), the graph representation of the image is updated. The superpixels within each cluster are merged into larger segments, as in the previous algorithms, but the segments features are recomputed and the similarity matrix \( W \) (with a reduced amount of vertices) is rebuilt. This updating scheme can be regarded as a “reset” to the simple agglomerative process, that takes place in different levels of the hierarchy. The reset reduces the amount of superpixels within the different clusters and thus the influence of this number on the linkage measure. As the updating procedure is not performed in each iteration of the algorithm, this scheme has the following benefits: 1) It uses the similarity between feature distributions of larger segments, as compared to the original similarity between neighboring superpixels. Such similarity relies on a coarser scale of information; 2) In part of the iterations merging is performed based on the linkage measure, thus its balancing effect on clusters size is maintained.
Chapter 4

Experimental results

This chapter presents experimental results for different steps within the proposed framework. Several image sets (Table 4.1) are used in the analysis. The cervigrams within these sets were randomly selected out of the NIH database, without any restricting rules. The images were digitized at 1660 dpi and compressed with 40:1 JPEG. Boundaries were marked and labeled by medical experts, with specialized experience in gynecological oncology, with the NLM Boundary Marking Tool (Jeronimo et al. [38]).

In $Set_1$ and $Set_2$ boundaries were marked by a single medical expert. $Set_3$ includes cervigrams in which two experts (not necessarily the same experts across all of the images) have marked the cervix boundary. $Set_4$ includes cervigrams that were automatically selected out of a larger set, in which the AW tissue was marked by two experts. These cervigrams were selected using the entropy-scaled-by-mean (ESM) segmentation complexity measure, presented in Chapter 5. The AW regions in the selected cervigrams are clearly visible and there is a large agreement among the experts in their markings.

<table>
<thead>
<tr>
<th></th>
<th>$Set_1$</th>
<th>$Set_2$</th>
<th>$Set_3$</th>
<th>$Set_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td># images</td>
<td>118</td>
<td>158</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td># experts</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>marked regions</td>
<td>cervix</td>
<td>cervix</td>
<td>cervix</td>
<td>AW</td>
</tr>
<tr>
<td></td>
<td>AW,CE</td>
<td>AW,CE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Sub-sets within the cervigrams archive

57
In order to quantitatively evaluate the segmentation results obtained in the following experiments, different performance measures are used. The following overlap measures are used when the segmentation quality of a single object (region / tissue) is addressed: the Dice measure: \( \frac{2|S \cap R|}{|S| + |R|} \), the Sensitivity measure: \( \frac{|S \cap R|}{|R|} \) and the False Positives (FP) measure: \( \frac{|S \cap \hat{R}|}{|\hat{R}|} \), with \( S \) being the area of the automatically segmented region, \( R \) the expert segmentation and \( \hat{R} \) its complement (the area outside the expert marked boundary). The Hausdorff distance, a maximum surface distance (Gerig et al. [21]), is used in part of the experiments. The Hausdorff distance defines the largest difference between two contours (measured in pixels) and is a good indicator for shape resemblance, with a smaller Hausdorff distance corresponding to more similar contours.

In order to represent the total segmentation quality (considering all tissues within the cervix region simultaneously) a variant of the Accuracy measure is used (Fawcett [18]). The Accuracy of a given classifier, defined on a binary set of samples with positive, \( P \), and negative, \( N \), labels, is computed as: \( \frac{TP + TN}{P + N} \), where \( TP \) and \( TN \) are the amount of true positives and true negatives detected by the classifier and \( P + N \) equals the sample size. When more than two labels are present, as in the current case of three tissues, Accuracy is defined as the sum of the true positives detected for each label (tissue), normalized by the sample (image) size. The higher the Accuracy, the better is the overall detection quality (more pixels are labeled correctly).

In order to accurately reflect the distribution of different results over the test set, notched box-and-whisker plots (McGill et al. [66]), hereon termed box-plots, are used throughout the different experiments. The box within this plot has lines at the lower quartile, median, and upper quartile values of the data. The whiskers are lines extending outward from the the box to show the extent of the rest of the data, with the extent set to 1.5 the interquartile range of the samples. Outliers are data with values beyond the ends of the whiskers (marked as +). The notches around the median line define a robust estimate of the uncertainty about the median and are used for box-to-box comparison. Two boxes whose notches do not overlap indicate that the medians of the two groups differ at the 5% significance level. When comparing between
different box-plots, the following rules are set in order to specify the best results: 1) The box-plot should have the highest (lowest) median, significantly different from the other medians; 2) The interquartile range (IQR) should be relatively narrow; 3) The values of the lower and upper quartiles should be as high (low) as possible.

Qualitative and quantitative results of the different steps within the automated analysis framework are provided next: Results for cervix boundary detection are presented in Section 4.1. Results for the illumination correction quality are presented in Section 4.2. Results for probabilistic pixel-based segmentation of cervix tissues are presented in section 4.3 and for tissues segmentation via clustering of superpixels are presented in Section 4.4. Section 4.5 presents preliminary classification results using two classification schemes along with a comparison to the probabilistic pixel-based segmentation. These results provide a summary for the current status of the cervix tissues detection quality.

4.1 Cervix boundary detection results

We start by evaluating the cervix boundary detection framework presented in the current work (Section 2.2). In the first experiment we compare the performance of the curve evolution process to clustering-based schemes for cervix-region extraction. One such scheme that performs clustering in the $\text{aR}$ feature space, was used for the initial ROI detection (Section 2.1.1). Several variations on the features used, feature normalization, and clustering techniques, were recently quantitatively compared by Xue et al. [97], using the same NIH cervigram test data ($\text{Set}_1$). The best results achieved in their experiments were obtained when the $\text{aR}$ features were linearly normalized to a unit variance and were clustered via Gaussian mixture modeling. The mean overlap results for this clustering-based variation are presented in Table 4.2 and compared to the mean overlap results of the curve-evolution framework (with the circular-prior method, Section 2.2.2-I). The results in Table 4.2 show a considerable improvement in the Dice and the FP measures using the curve evolution process. Sensitivity results slightly de-
crease, remaining at a strong level. These results indicate a more accurate delineation of the cervix region by the curve evolution process. Non-relevant tissues outside of the cervix region are better defined, which is an important step for further cervix analysis. A T-test is conducted in order to measure the significance of these results. The results of the clustering method are used as the null hypothesis for the test. The p-values computed for a 5% significance level are presented for each of the examined overlap measures. For \( p < 0.05 \), the results are stated as significantly different.

<table>
<thead>
<tr>
<th></th>
<th>clustering mean</th>
<th>curve evolution mean</th>
<th>curve evolution std</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>0.41</td>
<td>0.81</td>
<td>0.1</td>
<td>~0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1</td>
<td>0.94</td>
<td>0.08</td>
<td>~0</td>
</tr>
<tr>
<td>FP</td>
<td>0.35</td>
<td>0.13</td>
<td>0.12</td>
<td>~0</td>
</tr>
</tbody>
</table>

Table 4.2: Quantitative results for boundary detection: clustering-based schemes [97] vs the curve-evolution procedure with a circular shape prior. Mean results are presented along with the p-values computed for a 5% significance level (Set\(_1\)).

The effectiveness of the prior shape model for refinement of the initial data-driven segmentation (Section 2.2.1) in the curve evolution process, is explored next. The circular-prior method is examined in this case using both Set\(_1\) and Set\(_2\). A set of five cervigrams, randomly selected out of Set\(_1\), are used for the tuning of algorithm parameters. In the first step of the curve evolution process we set the parameters of Equation (2.9) to: \( \beta_1 = -3; \beta_2 = -20; \beta_3 = 20; \beta_4 = 0.01 \). The time step was set to: \( \Delta t = 2.5 \). The addition of the shape term in the second step requires a modification of the weights of the curvature-based forces within the data term to: \( \beta_2 = -10; \beta_3 = 10 \). Number of iterations in the second step were limited to 20. We found that these modifications were necessary to prevent the curve from shrinking into the cervix region in cases where its initial position is located within smooth regions and is already close to the desired result. The \( \gamma \) parameter in Equation (2.20) was set to \( \gamma = 4 \). It is important to note that once defined, the parameters were kept constant across all images in both test sets.

Figure 4.1 displays examples of cervix boundary detection. The data-driven contour
generated in the first step of the curve evolution process is marked in green and the
final result, following the refinement with the prior shape model, is marked in red.
Manual markings of the expert are shown in blue. Quantitative results are listed under
each image example, using overlap measures, along with the Hausdorff distance. Note
that the images are of approximately $1,500 \times 2,500$ pixels.

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>479</td>
<td>113</td>
</tr>
<tr>
<td>$D$</td>
<td>0.74</td>
<td>0.88</td>
</tr>
<tr>
<td>$S$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.24</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>226</td>
<td>74</td>
</tr>
<tr>
<td>$D$</td>
<td>0.87</td>
<td>0.94</td>
</tr>
<tr>
<td>$S$</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.17</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>245</td>
<td>176</td>
</tr>
<tr>
<td>$D$</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>$S$</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.14</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>216</td>
<td>93</td>
</tr>
<tr>
<td>$D$</td>
<td>0.83</td>
<td>0.93</td>
</tr>
<tr>
<td>$S$</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.26</td>
<td>0.007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>103</td>
<td>104</td>
</tr>
<tr>
<td>$D$</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>$S$</td>
<td>0.87</td>
<td>0.97</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.03</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>279</td>
<td>421</td>
</tr>
<tr>
<td>$D$</td>
<td>0.87</td>
<td>0.72</td>
</tr>
<tr>
<td>$S$</td>
<td>0.77</td>
<td>0.56</td>
</tr>
<tr>
<td>$FP$</td>
<td>0.14</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.1: Cervix boundary detection. Data-driven contour, based on color and cur-
vature features - marked in green; Final contour, following refinement with a circular
prior - marked in red; Expert markings imposed in blue. Hausdorff ($H$), Dice ($D$),
Sensitivity ($S$) and False Positives ($FP$), for the data-driven (data) and the circular
shape prior (shape) contours, are listed under corresponding cervigrams.

Several observations can be made: The contours are smoother and more convex
when the shape term is added, as desired. They exclude more irrelevant regions within
the cervix. Good similarity to the expert markings can be seen in most of the cases. These facts are reinforced by the quantitative evaluation, where a significant improvement in the FP and the Hausdorff distance is achieved. The sensitivity results of examples (a)-(e) attain high values, as they include most of the expert’s markings. These sensitivity results are similar for the two contours being compared. Example (f) shows an exceptional case, where the sensitivity decreases due to the shape prior constraint. In this example the data-driven curve is already close to, or located within the markings of the expert. Advancing the curve further using the circular shape prior generates a final smaller contour, as there are no color or curvature features that can prevent the curve from shrinking.

<table>
<thead>
<tr>
<th></th>
<th>Set₁</th>
<th></th>
<th>Set₂</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No prior</td>
<td>Circular</td>
<td>p-value</td>
<td>No prior</td>
</tr>
<tr>
<td>Dice</td>
<td>0.79(0.10)</td>
<td>0.81(0.10)</td>
<td>0.097</td>
<td>0.73(0.12)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95(0.05)</td>
<td>0.94(0.08)</td>
<td>0.076</td>
<td>0.96(0.07)</td>
</tr>
<tr>
<td>FP</td>
<td>0.28(0.13)</td>
<td>0.15(0.12)</td>
<td>~0</td>
<td>0.33(0.15)</td>
</tr>
<tr>
<td>Hausdorff</td>
<td>254(105)</td>
<td>216(91)</td>
<td>~0</td>
<td>300(125)</td>
</tr>
</tbody>
</table>

Table 4.3: Average results of cervix boundary detection with and without the circular shape prior. Mean value and standard deviation (within parentheses) are presented over the two data-sets used for evaluation (Set₁, Set₂). P-values computed for a 5% significance level are shown.

Table 4.3 summarizes average results of cervix boundary detection for Set₁ and Set₂ images, with (Circular) and without (No-prior) the shape priors. A paired T-test is used to evaluate if the results of the two methods are significantly different. P-values computed for a 5% significance level are shown. The Dice and Sensitivity values of the two methods are not significantly different. These results indicate that the data-driven curve is already in good proximity to the desired contour and that regions within the cervix are handled well. The Hausdorff distance shows a substantial improvement when adding the circular prior, thus indicating a better shape resemblance of the generated contours to the expert’s markings. The FP measure shows a considerable improvement, corresponding to a strong reduction of non-relevant tissues. This reinforces the addition of the shape term to the curve evolution process. Similar results are achieved for both
The next experiment examines the performance of the elliptical-prior method (Section 2.2.2-II). Tuning the different parameters within the active contour framework was performed on a small training set of five cervigrams, as in the circular-prior case. The same set of parameters, as in the previous experiment, was used for the data-driven curve in the first step of the curve evolution process. The addition of the elliptical prior in the second step required turning off the influence of the weighted region term. The $\gamma$ parameter in Equation (2.20) was set to $\gamma = 6$ and the number of iterations in the second step was limited to 10.

Figure 4.2: Cervix boundary detection with different shape priors. Manual markings imposed in blue; circular-prior results imposed in red and elliptical-prior results, imposed in white.

Figure 4.2 shows example results of the two shape priors. Manual markings are imposed in blue, results of the circular-prior are imposed in red and results of the elliptical-prior are imposed in white. Examples (a) and (b) show cases where cervix boundary detection by the circular-prior is more accurate. In these cases there is a high visual resemblance of the manual markings to a circle and the circular model is more appropriate. The cervix boundary in examples (c) and (d) has a high resemblance to an ellipse. The elliptical-prior achieved better results in these cases. Average results of the elliptical-prior, computed over the images of Set_1 and Set_2, are presented in Table
4.4 (Elliptical). The elliptical-prior results are compared to the circular-prior results using a paired T-test. The results of the two methods are not significantly different as indicated by the p-values. The results computed over the two data-sets are similar.

<table>
<thead>
<tr>
<th></th>
<th>Set(_1)</th>
<th></th>
<th>Set(_2)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Circular</td>
<td>Elliptical</td>
<td>p-value</td>
<td>Circular</td>
</tr>
<tr>
<td>Dice</td>
<td>0.81(0.10)</td>
<td>0.82(0.09)</td>
<td>0.19</td>
<td>0.75(0.13)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.94(0.08)</td>
<td>0.92(0.09)</td>
<td>0.19</td>
<td>0.95(0.09)</td>
</tr>
<tr>
<td>FP</td>
<td>0.13(0.12)</td>
<td>0.11(0.1)</td>
<td>0.45</td>
<td>0.16(0.14)</td>
</tr>
<tr>
<td>Hausdorff</td>
<td>216(91)</td>
<td>210(87)</td>
<td>0.15</td>
<td>231(115)</td>
</tr>
</tbody>
</table>

Table 4.4: Average results of cervix boundary detection with circular and elliptical shape priors. Mean value and standard deviation (within parentheses) are presented over the two data-sets used for evaluation (Set\(_1\), Set\(_2\)). P-values computed for a 5% significance level are shown.

In order to pursue a more thorough investigation of the two shape priors, we try to isolate the influence of the prior term on the evolving curve and reduce the effect of other factors that might interfere with the curve evolution process. Two special sub-sets of cervigrams were generated for this experiment, each containing 31 images. The first subset includes cervigrams in which the manual expert markings resemble a circle (the circular sub-set). The second sub-set includes cervigrams in which the markings resemble an ellipse (the elliptical sub-set). The data-driven curve obtained in the selected images, is already in good proximity to the manual markings of the expert, thus a good initialization for the second step of the curve evolution process is enabled. This good initialization prevents the curve from converging into local optimum, which are related to the image features and are located outside the actual cervix boundaries. The curve evolution process, in the selected cases, is mainly driven by the shape prior term, as desired.

Figure 4.3 presents box-plot results obtained with the circular-prior (left) and the elliptical-prior (right), in each of the sub-sets. The results for the circular sub-set are presented in the top row and for the elliptical sub-set are presented in the bottom row. Columns (I)-(IV) show results for: Dice, Sensitivity, FP and Hausdorff distance, in respective order. Table 4.5 presents the p-values computed for a paired T-test that
compares the circular-prior and elliptical-prior results of this experiment.

The results of the two shape priors on the circular sub-set are not significantly different as shown both from the box-plots and from the computed p-values. This observation indicates that the elliptical-prior can capture well a circular cervix region, as the ellipse is a generalization of a circle. In the results of the elliptical sub-set, the elliptical-prior obtained significantly better results, as expected.

<table>
<thead>
<tr>
<th></th>
<th>Circular sub-set</th>
<th>Elliptical sub-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>0.58</td>
<td>≈0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.86</td>
<td>0.25</td>
</tr>
<tr>
<td>FP</td>
<td>0.18</td>
<td>≈0</td>
</tr>
<tr>
<td>Hausdorff</td>
<td>0.65</td>
<td>≈0</td>
</tr>
</tbody>
</table>

Table 4.5: A comparison between the circular-prior and the elliptical-prior on the circular sub-set and the elliptical sub-set using a paired T-test. P-values computed for a 5% significance level are shown.

In the final experiment the results of the suggested framework (with the circular
prior) were compared to the markings of two experts, as available in Set$_3$. Figure 4.4 presents several example segmentations. The manual markings of the two experts are shown in green and blue. The results of the algorithm are imposed in red. The Dice and the Sensitivity measures, computed between the two experts and between the algorithm and each of the experts, are listed below corresponding images. The sensitivity between the experts was computed twice per image and averaged, using each of the experts as the ground truth.

<table>
<thead>
<tr>
<th></th>
<th>Dice$_e$</th>
<th>Sens$_e$</th>
<th>Dice$_1$</th>
<th>Sens$_1$</th>
<th>Dice$_2$</th>
<th>Sens$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>0.79</td>
<td>0.82</td>
<td>0.84</td>
<td>0.73</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>(b)</td>
<td>0.90</td>
<td>0.91</td>
<td>0.85</td>
<td>0.94</td>
<td>0.90</td>
<td>0.94</td>
</tr>
<tr>
<td>(c)</td>
<td>0.89</td>
<td>0.90</td>
<td>0.94</td>
<td>0.89</td>
<td>0.84</td>
<td>0.73</td>
</tr>
<tr>
<td>(d)</td>
<td>0.51</td>
<td>0.67</td>
<td>0.50</td>
<td>1.00</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>(e)</td>
<td>0.45</td>
<td>0.65</td>
<td>0.76</td>
<td>0.62</td>
<td>0.62</td>
<td>0.99</td>
</tr>
<tr>
<td>(f)</td>
<td>0.55</td>
<td>0.69</td>
<td>0.65</td>
<td>1.00</td>
<td>0.87</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Figure 4.4: Automated boundary detection (red) as compared to the markings of two experts (green and blue). Listed are the corresponding Dice and Sensitivity measures between the two experts ($\text{Dice}_e, \text{Sens}_e$) and between the algorithm and expert 1 and expert 2 ($\text{Dice}_1, \text{Sens}_1, \text{Dice}_2, \text{Sens}_2$, respectively).

Cervigrams (a)-(c) are examples of strong agreement between the expert markings. In these cases the automated cervix delineation results are consistent with these of the experts. In examples (d)-(f) there is a strong disagreement between the expert markings of the cervix region. The algorithm results are close to one of the experts, or
within the range of their markings. Table 4.6 summarizes the results for \( Set_3 \) images. The average Dice result between the two experts is 0.88. The standard deviation of 0.1 indicates a strong variability between the experts. Similar results are obtained for the Sensitivity measure. The results of the experts are more accurate than those of the algorithm and are significantly better (as indicated by the p-values), but the overlap that exists between the two distributions indicates that there are many cases in which the algorithm has marked the images in a similar way to at least one of the experts.

<table>
<thead>
<tr>
<th></th>
<th>expert1 vs expert2</th>
<th>algorithm vs experts</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>mean</td>
</tr>
<tr>
<td>Dice</td>
<td>0.88</td>
<td>0.1</td>
<td>0.81</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.9</td>
<td>0.07</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4.6: Average Dice and Sensitivity measures for a comparison between the algorithm and the results of two experts (\( Set_3 \)). P-values computed for a 5% significance level are shown.

### 4.2 Illumination correction results

In this section the need for an illumination correction process to improve cervix tissues segmentation is examined and the quality of the GEM-based illumination correction (Section 2.3) is evaluated by a comparison to the state-of-the-art method of MultiScale Retinex (MSR) (Rahman et al. [76]). Segmentation via an unsupervised modeling scheme (based on a mixture of three Gaussians in \( Lb \) feature space (Section 3.1.2)) is used for this purpose. The experiments were performed on \( Set_1 \). The validation was conducted within the cervix region as marked by the expert in order to ignore errors of the automatic cervix boundary detection. SR pixels are ignored in the analysis.

Figure 4.5 exemplifies four segmentation results, before and following the illumination correction by the GEM and the MSR algorithms. The top row of each example includes (from left to right): (1) The original cropped cervigram with manual markings of the expert imposed. AW regions are marked by a blue line, CE regions are marked by a purple line and the SE region is defined as the rest of the tissue within
the cervix boundary, which is marked by a yellow line; (2) The original cropped image with the SR and non-cervix regions masked out (marked in black); (3), (4) The illumination corrected images by the MSR and the GEM algorithms, respectively; (5) The illumination field estimated by the GEM algorithm. The bottom row of each example presents the corresponding segmentation maps as marked by the expert (1) and by the unsupervised segmentation (2, 3 and 4). Within these maps, AW regions are colored dark-red, CE regions are colored light blue and SE regions are yellow. The dark blue regions are ignored throughout the evaluation process. They correspond to other manually marked tissues, to SR pixels, or to regions outside the cervix boundary (this color convention is used throughout the rest of the experiments). The Accuracy measure computed for each case is listed under corresponding segmentation maps.

Visual inspection of the estimated illumination (column (5)) reveals the varying illumination fields across the cervigrams. A correspondence can be seen between the estimated illumination field and the detected AW regions in the segmentation maps of the original images (column (2)). Falsely detected AW regions are reduced following the illumination correction step by both the MSR (column (3)) and the GEM (column (4)) algorithms. In the segmentation results following the MSR algorithm, large portions of falsely detected AW regions still remain. A better result can be seen following the GEM algorithm, where the CE segmentation is also better. These visual observations are reinforced by the values of the Accuracy measure, which are the highest following the GEM-based illumination correction.

Figure 4.6 presents box-plot results of the segmentation quality obtained in this test. Segmentation using unsupervised modeling (US) was used throughout the different cases. Results for segmentation of the original images (US-ORG) and following the MSR (US-MSR) and GEM-based (US-GEM) illumination correction, are presented. The figure shows results of the Accuracy measure, for a combined evaluation as well as of the Dice measure, computed for each individual tissue. The following observations can be made: 1) The medians of the Accuracy and $SE_{\text{Dice}}$ of the US-GEM method, are significantly better (shown by the notch) as compared to the other methods. This
Figure 4.5: Illumination correction examples: (1) Original cervigram, expert markings superimposed; (2) Preprocessed cervigram, specularities and non cervix regions are masked out; (3),(4) Illumination corrected cervigrams by MSR and GEM, respectively; (5) Illumination field estimated by GEM. Corresponding unsupervised segmentation maps including Accuracy results, are presented under each image. AW regions are colored dark-red, CE regions are light-blue and SE regions are yellow. SR and non-cervix regions are colored dark-blue.
observation indicates that the SE is more accurately segmented, thus a reduced amount of false positives from the other tissues is present in the segmentation; 2) The median of the $AW_{Dice}$ result of the US-GEM method is slightly better than the median of the US-ORG method, the upper quartile is slightly damaged; 3) The medians of the US-GEM case are significantly better than those of the US-MSR for the different quality measures and the position of the interquartile range is higher. This observation indicates that segmentation following illumination correction via GEM is better than following correction via MSR; 4) A wide IQR can be seen in all of the examined box-plots, thus indicating a large variability in the results. This observation indicates that although the illumination correction supports the segmentation process, the segmentation results needs to be further improved.
4.3 Analysis of probabilistic pixel-based cervix tissues segmentation

We next present a feasibility study for the probabilistic pixel-based tissue segmentation framework (Section 3.1). The MAP-EM algorithm is utilized in order to generate a range of tissue models for that purpose. Starting from $\beta = 0$ (unsupervised modeling) to $\beta \to \infty$ (supervised modeling), varying $\beta$ values are used to control the relative influence of the prior tissues model on the observed data of the input image. The $\beta$ constant is defined in the current experiment as a percentage from the number of samples in the input image (it is multiplied by this number during the learning phase). This is done in order to maintain a similar influence of the prior model (per $\beta$ value) across images with different size. Increasing percentage values of $\beta = [0, 0.2, 0.5, 1, 5, 10, 100, \infty]$ are used.

We use a set of 35 manually marked cervigrams that were randomly selected out of Set$_1$ to learn the prior tissue models. A similar amount of samples is extracted per tissue out of the different tissue markings across the images and is used as a training set for the tissue modeling. Each tissue is next modeled as a Gaussian in $Lb$ feature space based on this training set.

Parameter initialization is an important issue in EM-based algorithms due to their tendency to converge to a local minimum. Here the mixture of Gaussians (MoG) parameters were initialized using a K-means clustering procedure in the unsupervised modeling case and the unsupervised MoG parameters were used to initialize the MAP-EM process for the different $\beta$ values. The experiment was performed on Set$_1$. The validation was conducted within the cervix region as marked by the expert. SR pixels are ignored in the analysis.

Figure 4.7 presents box-plot results of the Accuracy and Dice values computed in this experiment over the entire image-set. The $X$-axis corresponds to the increasing $\beta$ values: the left-most point corresponds to segmentation via unsupervised modeling and the right-most point corresponds to segmentation via supervised modeling. From the
presented box-plots, two main observations are made: 1) The best Accuracy results, along with $SE_{\text{Dice}}$ and $CE_{\text{Dice}}$ results, are obtained for $\beta = 0.5$ (box-plot 3) and $\beta = 1$ (box-plot 4), as the medians and IQR results of these cases are higher. This indicates that for the entire tissue set, as well as for these tissues, best results are achieved with equal contribution from the prior models and the observed data; 2) $AW_{\text{Dice}}$ results are improved as the $\beta$ value is increased. Thus, in the AW tissue case, results using supervised modeling (box-plot 8) are significantly better than these of the unsupervised modeling case (box-plot 1).

Figure 4.7: Box-plot results of the MAP-EM process for different $\beta$ values. X-axis corresponds to $\beta = [0, 0.2, 0.5, 1, 5, 10, 100, \inf]$, in increasing order ($Set_1$).

Figure 4.8 presents some example segmentation results using the MAP-EM framework. The original image with the expert markings imposed is presented in column (1); the corresponding manual segmentation maps are presented in column (2); Segmentation maps for $\beta = 0$ (unsupervised modeling), $\beta = 1$, and $\beta = \inf$ (supervised modeling) are presented in columns (3)-(5), respectively. Computed Accuracy (left) and $AW_{\text{Dice}}$ (right) values are listed for each result. These examples demonstrate the behavior of the different quality measures, as indicated by the box-plots of Figure 4.7. The Accuracy measure for $\beta = 1$ (column(4)) is higher than in the unsupervised case.
Figure 4.8: MAP-EM segmentation results for different $\beta$ values. (1) Expert markings; (2) Corresponding manual segmentation maps; (3),(4),(5) Segmentation maps for $\beta = [0, 1, \infty]$. Accuracy (left) and $AW_{Dice}$ (right) measures are listed under corresponding results. AW regions are colored dark-red, CE regions are light-blue and SE regions are yellow. SR and non-cervix regions are colored dark-blue.
(column(3)) and better or similar to the accuracy of the supervised case (column(5)).
The amount of falsely detected tissues (both for AW and CE) is better in that case, as
can be seen from the corresponding segmentation maps. With regards to AW segment-
tation, more pixels are detected correctly in the maps of the supervised case (increased
$AW_{Dice}$ results). Especially interesting is example IV, where no AW region exists (thus
$AW_{Dice} = 0$). A very small amount of falsely detected AW pixels is present in this case
for $beta = 1$ and $\beta = \inf$. This example demonstrates the importance of using a prior
model in such cases.

![Segmentation Results](image)

Figure 4.9: Summary of segmentation results via unsupervised (US) and supervised
(SUP) modeling, on original images (ORG) versus illumination corrected ones (GEM)
(110 images).

Figure 4.9 presents a summary of the main results achieved with the pixel-based
segmentation framework, by comparing unsupervised (US) and supervised (SUP) mod-
eling on both the original images (ORG) and the illumination corrected (GEM) images.
Box-plots of Accuracy and Dice measures are used. P-values (computed with a 5% sig-
nificance level) for a paired T-test that compares between the different methods, are
presented in Table 4.7. The Accuracy results following the illumination correction pro-
cess, when using the supervised modeling scheme (SUP-GEM) are significantly better
than the other examined methods. The median of the $AW_{Dice}$ results of the SUP-GEM framework, although still at a low value, is significantly better than the medians of the other methods.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>$SE_{Dice}$</th>
<th>$CE_{Dice}$</th>
<th>$AW_{Dice}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-GEM vs. SUP-GEM</td>
<td>0.02</td>
<td>0.61</td>
<td>0.66</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>SUP-ORG vs. SUP-GEM</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>0.58</td>
<td>$\sim 0$</td>
</tr>
</tbody>
</table>

Table 4.7: P-values computed for the different comparisons presented in Figure 4.9 between segmentation via unsupervised (US) and supervised (SUP) modeling, on original images (ORG) versus illumination corrected ones (GEM). P-values were computed for a 5% significance level.

4.4 Analysis of segmentation via clustering of superpixels

The next set of experiments is focused on the different steps within the cervix tissues segmentation framework via clustering of superpixels (Section 3.2). Experiments were performed on $Set_1$ and on $Set_4$. The analysis on $Set_1$ is performed within the automatically detected cervix boundary (using the data-driven curve evolution process, Section 2.2.1). The analysis in $Set_4$ is performed within the cervix boundary as marked by the medical experts. The results on $Set_4$ are expected to be better as the cervix boundaries are more accurate and the AW lesions, which are clearly visible, are marked by the experts in a similar way. The images within the two sets were preprocessed for SR detection and illumination correction.

The main objective of this framework is to automatically generate segments, which are homogeneous regions of similar content and of any shape, that make up the tissue regions. Once such segments are extracted, a classification phase is required in order to associate the segments with corresponding tissue labels. The main assumption of the current step is that a better overlap of segments with the cervix tissues, will provide better results in a future classification step.

Evaluation of the work entails a measure of overlap between a mask, generated
per AW or CE tissue, and the manual markings of the medical expert. The mask is
built out of an ensemble of segments, which are covered by the region of the expert-
marked tissue of interest. The segments are associated with the corresponding tissue
mask using the following majority rules: 1) Associate the segment with the tissue that
covers most of its area; 2) If the segment is covered by all tissue types: AW, CE, SE and
the SE region occupies less than 50% of its area, associate the segment with the largest
among the AW or CE regions. Following the generation of the AW and CE masks
(hereon referred to as AW and CE segmentation results), different quality measures
are used to quantify the overlap between the automated and the expert-based tissue
delineation.

Figure 4.10 presents example segmentations that demonstrate the ability of the
current framework to capture regions of non-convex nature within the cervix. Original
cervigrams are presented in the top row, where AW regions are marked in blue by a
medical expert and the cervix boundary is marked in yellow or green. Corresponding
segmentation results are presented in the bottom row, where segments boundaries are
outlined in black (regions outside the cervix are masked out). These segments were
generated by the Ag clustering algorithm and the NMCut criterion. The hierarchy
level of ten segments is shown. AW Dice, Sensitivity and FP values are listed above
corresponding segmentations. The elongated and non-convex nature of the generated
segments is clearly visible within these examples. The AW segmentation quality is high
as reflected by the values of the different overlap measures.

The following experiments perform a thorough analysis of the different steps within
the presented framework. We first examine the quality of the generated superpixels
in Section 4.4.1. The ability of the local scaling process to increase the differences
between inter and intra-tissue boundary portions is investigated in Section 4.4.2. The
influence of the similarity matrix on segmentation results is examined in this section as
well. Finally the different agglomerative clustering algorithms and graph-cut criterions
are evaluated in Section 4.4.3.
Figure 4.10: Segmentation examples. Top row: ground truth markings on original cervigrams. AW marked in blue. Cervix region marked in yellow or green. Bottom row: segmentation results for Ag clustering with the NMCut criterion for ten segments. Segments boundaries are marked in black. Regions outside the cervix are masked out. AW Dice / Sensitivity / FP values are listed above corresponding segmentations.

4.4.1 Quality analysis of superpixels overlap with cervix tissues

The superpixels, which are the basic components of the current framework, may introduce their own errors when compared to the manual markings of the expert. The current experiment measures the overlap between an ensemble of superpixels, collected per tissue and the manual tissue markings. This overlap is regarded as the reference ground truth (GT), as it indicates the best results that can be achieved by merging the superpixels correctly.

Figure 4.11: Superpixels accuracy
Figure 4.11 presents box-plot results of the Dice, Sensitivity and FP overlap measures, obtained in this experiment. The box-plots within each figure present overlap results for the CE tissue in Set_1 (left box-plot), the AW tissue in Set_1 (middle box-plot) and the AW tissue in Set_4 (right box-plot). From the results on Set_1 we can observe that the overlap between the superpixels and the CE tissue is better than the overlap between the superpixels and the AW tissue (the median Dice is 0.9 for CE and 0.87 for AW). The overlap between the superpixels and the AW tissue in Set_4 is much better (median Dice is 0.94), as expected. These results indicate that a correct ensemble of superpixels covers the tissues of interest with an acceptable accuracy and that the superpixels generated in this step can be used as an initialization for the successive merging process.

4.4.2 Quality analysis of the similarity matrix

In the following experiments we assess the quality of the similarity matrix suggested in Section 3.2.2 and compare it to the quality of other available similarity matrices. We start by evaluating the quality of the local scaling process of the edge cost, suggested in Section 3.2.2-II. In order to obtain good clustering results, weights should be high for intra-tissue boundaries and low for inter-tissue boundaries. The absolute difference between these two groups serves as a quality measure. The analysis is performed with respect to the tissue boundaries as marked by a medical expert. The experiment is performed on a set of 100 images from Set_1 for which the manual markings of the AW and the CE tissues are available (the cervigrams contain at least one of these tissues).

The quality measure for comparison, is computed in the following manner (as illustrated in Figure 4.12): Given a similarity matrix $W$, the weights $w_{ij}$ are imposed on their respective boundary portions in the map of superpixels (Figure 4.12(c)). Note the high weights of boundary portions between similar superpixels in the figure, which are color-coded with the yellow-red shades of the scale. Boundary portions between different superpixels obtain low weights and are color-coded with the blue shades of the scale. The map is next compared with a ground truth (GT) edge map, which is extracted
out of the different boundary portions, to best feet the manual markings of the expert (Figure 4.12(d)). This map is in good proximity to the manual markings of the tissues within the original cervigram (Figure 4.12(a)). The boundary portions are divided into two groups: one group includes the inter-tissue boundary portions, located on the GT boundaries (Figure 4.12(e)). The second group includes the intra-tissue boundary portions, located within the tissues regions (Figure 4.12(f)). The mean values of the boundary portions within the two groups: \( \text{mean}_{\text{inter}} \) and \( \text{mean}_{\text{intra}} \) are computed next and the absolute difference between these two means: \( \text{mean}_{\text{diff}} \) is used as the quality measure for the given similarity matrix. The best similarity matrix is selected as the one with the highest \( \text{mean}_{\text{diff}} \) value. The example in Figure 4.12 illustrates the low similarity values obtained for most of the inter-tissue boundaries and the high similarity values obtained for most of the intra-tissue boundaries. This observation is reinforced by the mean values of the two groups: \( \text{mean}_{\text{inter}} = 0.31, \text{mean}_{\text{intra}} = 0.59 \).

Four local scaling schemes are compared. The first scheme, \( E^1 \), uses the original energy of the normalized gradients along the boundary portion: \( e_{ij} \), without scaling. The second scheme, termed \( E^2 \), includes a local scaling process, which is based on the morphological Toggle-Contrast transform (Soille [85]). This process is defined for graphs via the following equation:

\[
E^2_{ij} = \begin{cases} 
  e_{\text{max}} & \text{if } e_{\text{max}} - e_{ij} < e_{ij} - e_{\text{min}} \\
  e_{\text{min}} & \text{otherwise}, 
\end{cases}
\]  

\( (4.1) \)

where \( e_{\text{max}} = \max_{(k,l)\in N}(e_{kl}) \), \( e_{\text{min}} = \min_{(k,l)\in N}(e_{kl}) \) and \( (k,l) \in N \) is the boundary portions neighborhood (boundary portions connected on both ends of boundary \( (i,j) \)), including boundary \( (i,j) \)). The third scheme, \( E^3 \), uses a local scaling process which is based on a sigmoid kernel and is computed as:

\[
E^3_{ij} = \begin{cases} 
  e_{ij} & \text{if } e_{\sigma} < T \\
  \sqrt{\frac{e_{ij}}{1+\exp\left(-\frac{e_{ij} - e_{\mu}}{e_{\sigma}}\right)}} & \text{otherwise}, 
\end{cases}
\]  

\( (4.2) \)

where \( e_{\mu} \) and \( e_{\sigma} \) are the mean and standard deviation of the boundary portions neigh-
Figure 4.12: Similarity matrix quality computation: (a) Original cervigram. AW boundary imposed in blue. Cervix boundary imposed in yellow; (b) Superpixels decomposition, boundaries imposed in black. Regions outside the automatically detected cervix region are masked out (black); (c) Weights of similarity matrix imposed on corresponding boundary portions; Blue shades correspond to weak similarities and yellow-red shades to strong similarities; (d) GT edge map used for computation; (e) Inter-tissue boundaries; (f) Intra-tissue boundaries.
neighborhood. The thresholding condition is used to limit unnecessary scaling (same as in Equation (3.7)). The threshold is set to $T = 0.005$ in this experiment. The last scheme, $E^4$, is the local scaling process presented in Section 3.2.2-II. Figure 4.13 presents box-plot results of the $\text{mean}_{\text{dist}}$ quality measure computed for each of the examined local scaling schemes. Box-plot results for schemes $E^1 - E^4$, are presented from left to right respectively. The best results are obtained for scheme $E^4$, which is used in the current framework. The median of the results is significantly better in this case as compared to the other local scaling schemes and to no scaling at all. P-values (computed with a 5% significance level) of a paired t-test that compares between scheme $E^4$ and each of the other schemes are $\sim 0$.

![Figure 4.13: A comparison between different local scaling schemes of the edge cost. Box-plot results for schemes $E^1 - E^4$, are presented from left to right respectively.](image)

We next evaluate the influence of the similarity matrix on the agglomerative clustering process itself. Three types of similarity matrices are compared. Within these matrices, superpixels are represented in $Lb$ feature space. Edge cost scheme $E^4$ is used for the final similarity computation. Three schemes for region similarity are used: 1) A linearly normalized similarity matrix based on a $d$-dimensional histogram (with 25 bins per channel) for the superpixels representation. Region similarity between superpixels is computed in this case via the symmetric version of the discrete KL divergence, where the KL divergence between two histograms, $p$ and $q$, with $n$ bins, is defined as: $\sum_{k=1}^{n} p_k \log \frac{p_k}{q_k}$. This option is hereon termed the normalized-histogram-
KL ($nhKL$); 2) A locally scaled version of the histogram-based similarity matrix of the $nhKL$ scheme, where the matrix is locally scaled with the self tuning kernel of Equation (3.5). This scheme is hereon termed the scaled-histogram-KL ($shKL$); 3) A locally scaled similarity matrix using a Gaussian representation per superpixel and the KL divergence between Gaussians, as presented in Section 3.2.2-I. This scheme is hereon termed the scaled-KL ($sKL$).

Figure 4.14 presents the AW (left) and CE (right) segmentation results obtained in this experiment in the last 30 steps of the agglomerative clustering process. The $X$-axis in each plot corresponds to the number of segments present in the current step. The figure shows average Dice and Sensitivity results, computed per step, for $Set_1$. The $nhKL$ results are plotted in a dotted line, the $shKL$ results are plotted in a dashed line and the $sKL$ results are plotted in a solid line. Several observations can be made from the plots: AW segmentation results are very similar for both the $shKL$ and the $sKL$ matrices. CE segmentation results with the $sKL$ matrix are better as the number of segments decrease. In both tissue types the results of the $nhKL$ matrix are inferior to these of the locally scaled matrices. The FP values obtained in these experiments are very low and non-informative and are therefore omitted from the plots. This result is due to the way the AW and CE masks are generated. The masks are built from an ensemble of segments, in which the corresponding tissue has a majority of pixels, thus the amount of FP is relatively low by definition. The FP results are hereon omitted from the analysis.

Finally we examine the distribution of the weights within the $sKL$ similarity matrix (computed per image) over the cervigrams of $Set_1$. Figure 4.15 presents two overlapped histograms, one for the collection of inter-tissue similarities extracted from each cervigram (black), and one for the collection of intra-tissue similarities (gray). From these histograms it can be seen that the weights in both groups occupy the full range of available values. The weights of the inter-tissue similarities are more concentrated at the lower part of the scale, as expected, but there is no clear distinction between the two groups. This result indicates the large variability that exists within the database.
A single threshold may not exist that can perfectly separate between the two groups. This fact may lead to poor automatic segmentation results over the entire image set.

![Figure 4.14: A comparison between different similarity matrices via agglomerative clustering. Average AW (left) and CE (right) segmentation results are presented per step. Results are presented for up to 30 segments (Set1).](image)

### 4.4.3 Quality analysis of the agglomerative clustering framework

In this section the agglomerative clustering framework is evaluated. The $sKL$ similarity matrix is used throughout the different experiments.

I. **A Comparison between different cut criterions via agglomerative clustering**

The simple agglomerative clustering algorithm (Ag) is used to compare between the different graph cuts criterions, described in Section 3.2.3-I: the weighted-mean cut
Figure 4.15: Distribution of inter-tissue (black) and intra-tissue (gray) similarities computed within cervigrams of Set1.

(WMCut), the MinMax-cut (MinMax) and the normalized-mean cut (NMCut), presented in this work. Figures 4.16-4.18 show example results for segments generated in different steps of the agglomerative clustering process, using the different graph-cut criterions. The original cropped image, with the expert tissue markings imposed, is shown in (a). The manually marked cervix regions is outlined in yellow, AW tissue is outlined in blue and CE tissue in purple. The automatically detected cervix boundary is outlined in white in (b). Superpixels boundaries are imposed in black on a preprocessed image in (c). The image is preprocessed for SR elimination and illumination correction. Regions outside the automatically detected cervix boundary are masked out. Segment boundaries generated in steps: 50, 30 and 10 of the agglomerative hierarchy, are shown in rows (1)-(3), respectively. The boundaries within each row were generated using the $\ell_{WMCut}$ (column (c)), $\ell_{MinMax}$ (column (d)) and $\ell_{NMCut}$ (column (e)) linkage measures. Corresponding Dice, sensitivity and FP measures computed for the AW tissue, are listed above each segmentation result.

Visual inspection of the superpixels within each figure (b), indicates that most of the superpixels do not cross any important tissue boundaries. In particular, an overlap with the expert marked AW boundaries can be detected. There are very few cases where this observation doesn’t hold. In these cases the boundary energy is weaker than the threshold specified during the superpixels generation process (Section 3.2.1). Such a
case can be detected in example III (Figure 4.18). In this example, the superpixel surrounding the AW lesion at the lower part of the cervix, is larger than the marked AW boundaries.

The quality of the segmentation results for the different graph cut criterions is shown to decrease as the number of segments is reduced. This is due to erroneous merging of segments, which is an outcome of the merging process and the linkage measure. The following observations are made from the segmentation maps with ten segments (row (3)): 1) The $\ell_{WMCut}$ provides large segments that cover large parts of the image plane and some small isolated segments. The AW region covers a small portion out of these large segments thus the segments are not considered part of the generated AW mask and the Dice and Sensitivity values are very low (zeroed); 2) Segments generated by the $\ell_{MinMax}$ criterion are of similar (balanced) size and convex in shape. The AW region is erroneously merged with the surrounding tissues in part of the cases; 3) The $\ell_{NMCut}$ criterion provides elongated segments, which are more balanced in size. These segments capture the non-convex shape of the tissues within the cervix more accurately.

Example III, is a case where the agglomerative framework fails. The size of the AW lesions in this example is relatively small as compared to the entire cervix region. Note the bottom AW lesion, which is contained within a single superpixel. In addition to size, the AW region and edge features in this example are not distinctive enough and even visual detection of the lesions is difficult. The generated segments in this case tend to merge with the surrounding superpixels during the agglomerative clustering process, as there is no limiting condition on clustering. Cases like this motivate the Constrained Agglomerative Clustering algorithm presented in Section 3.2.3-II.

Figure 4.19 summarizes segmentation results obtained in this experiment for Set$_{1}$ (top) and for Set$_{4}$ (bottom). The average Dice and Sensitivity results are shown for the last 50 steps of the agglomerative clustering process and the different tissues marked within each set. WMCut results are plotted with a dotted line, MinMax-cut results are plotted with a dashed line and NMCut results are plotted with a solid line. The following observations can be made: 1) Results at the step of 50 segments are similar for
(a) (b) (c)

0.84/0.92/0.10
0.91/0.92/0.04
0.91/0.93/0.05

(2)

0.73/0.73/0.11
0.88/0.83/0.02
0.83/0.92/0.12

(3)

0.02/0.01/0.00
0.71/0.65/0.07
0.74/0.74/0.10

(d) (e) (f)

Figure 4.16: Agglomerative segmentation example I: (a) Original image with expert markings imposed. Cervix region marked in yellow, AW marked in blue; (b) Automatically detected cervix boundary imposed in white; (c) Superpixels boundaries imposed in black on a preprocessed image; Regions outside the automatically detected cervix boundary are colored black. Rows (1)-(3) correspond to: 50, 30, 10 segments; Columns (d)-(f) correspond to: WMcut, MinMax-cut, NMcut respectively. AW Dice/Sensitivity/FP values are listed above corresponding segmentations.
Figure 4.17: Agglomerative segmentation example II: (a) Original image with expert markings imposed. Cervix region marked in yellow, AW marked in blue, CE in purple; (b) Automatically detected cervix boundary imposed in white; (c) Superpixels boundaries imposed in black on a preprocessed image; Regions outside the automatically detected cervix boundary are colored black. Rows (1)-(3) correspond to: 50, 30, 10 segments; Columns (d)-(f) correspond to: WMcut, MinMax-cut, NMcut respectively. AW Dice/Sensitivity/FP values are listed above corresponding segmentations.
Figure 4.18: Agglomerative segmentation example III: (a) Original image with expert markings imposed. Cervix region marked in yellow, AW marked in blue; (b) Automatically detected cervix boundary imposed in white; (c) Superpixels boundaries imposed in black on a preprocessed image; Regions outside the automatically detected cervix boundary are colored black. Rows (1)-(3) correspond to: 50, 30, 10 segments; Columns (d)-(f) correspond to: WMcut, MinMax-cut, NMcut respectively. AW Dice/Sensitivity/FP values are listed above corresponding segmentations.
all graph-cut criterions ($AW_{\text{Dice}} = 0.57$, $CE_{\text{Dice}} = 0.73$ for $Set_1$ and $AW_{\text{Dice}} = 0.89$ for $Set_4$). The Dice and Sensitivity values decrease along the agglomerative process due to erroneous merging. The best linkage measure is considered to be the one that obtains the best results throughout the different steps; 2) The WMCut obtains poor results: The quality of the AW segmentation results with this criterion decreases more rapidly as compared to the other criterions; 3) AW results of the MinMax-cut and the NMcut are not significantly different. The NMcut results are slightly better throughout most of the agglomerative steps in $Set_1$; 4) For the CE region, which is more convex in shape, all three graph-cut criterions obtain similar results; 5) The overall Dice and sensitivity results in $Set_4$ are better than the results in $Set_1$, as expected.

Figure 4.20 focuses on five steps within the agglomerative process presented in Figure 4.19. The figure presents box-plot results for the steps of 50, 40, 30, 20, 10 segments. The box-plots compare between the different graph-cuts: WMCut, MinMax-cut and NMcut, from left to right within each step (the results per step are separated by a vertical grid). We note the following: 1) The IQR range for the AW segmentation results in $Set_1$ is very large, which indicates that the algorithm doesn’t work well for all existing cases. There are cases where the generated masks don’t capture any superpixel, thus the results are zero; 2) The IQR of the box-plots within $Set_4$ are significantly narrower as compared to $Set_1$, this indicates on the more homogenous nature of $Set_4$; 3) The NMcut criterion obtains the best Dice and Sensitivity median values for AW segmentation in $Set_1$. This result becomes more significant as the number of segments decrease. On $Set_4$ the NMcut results are slightly better than these of the MinMax-cut; 4) The CE segmentation results of the NMcut and MinMax-cut are similar and better than these of the WMCut.

II. Constrained agglomerative clustering analysis

This experiment investigates the influence of the threshold $T$, within the stopping rule of the constrained agglomerative (CAg) algorithm (Section 3.2.3-II), on the final segmentation results. The threshold is applied to the weights between neighboring
Figure 4.19: Average Dice and Sensitivity segmentation results for agglomerative clustering with different graph-cut criterions. Results are presented per step, for up to 50 segments. Top: AW and CE segmentation results for $Set_1$; Bottom: AW segmentation results for $Set_4$. 

- AW
- CE
- $Set_1$
- $Set_4$
Figure 4.20: Box-plot results with agglomerative clustering, for 50, 40, 30, 20, 10 segments. A comparison across different graph-cut criterions: WMCut, MinMax-cut, NMCut, from left to right within each group, respectively. Top: AW and CE segmentation results for Set_1; Bottom: AW segmentation results for Set_4.
superpixels. The following thresholds: $T = 0.1, 0.2, 0.3, 0.4, 0.5$, are examined. The NMCut criterion is used for merging. Figure 4.21 presents box-plots that summarize the segmentation results obtained for cervigrams of $Set_1$ (top) and of $Set_4$ (bottom) for the different thresholds (X-axis), used in the experiment. Figure 4.22 presents box-plots of the final number of segments, generated per image, with each of the thresholds (for $Set_1$). The number of segments is shown to increase with the increase of the threshold. A large amount of clusters also improves the segmentation quality, as less erroneous merging are present. Thus the best results are obtained for $T = 0.5$. Note that the median number of segments in that case is around 60. AW results on $Set_4$ are better than on $Set_1$.

Figure 4.23 presents example results of the CAg algorithm on the same cervigrams from Figures 4.16-4.18. Segment boundaries generated during the CAg clustering process are shown for the thresholds of $T = 0.1, 0.3$ and 0.5. Dice, Sensitivity and FP measures computed for the AW tissue, are listed above corresponding segmentations. Visual inspection of the segmentation maps reveals that the maps generated with different thresholds are not contained within each other, as in the agglomerative case and each threshold leads to a different optimum. The increase in segments number is shown to be correlated with the increase of the threshold and with a better segmentation quality. The results with $T = 0.1$, where the order of 10 segments is present, are inferior when compared to the Ag results with 10 segments. The last example in this set, that was used to motivate the CAg process (example III, Figure 4.18), is shown to provide poor computational results. When visually inspecting the corresponding segments it can be seen that the AW region located in the upper part of the cervix is captured more accurately throughout the different thresholds. The generated segments (though not accurate enough) are better than the ones obtained with the Ag algorithm, where the AW region was forced to merge with segments of other tissues.
Figure 4.21: CAg-segmentation: Dice and Sensitivity box-plot results with different thresholds (X-axis). Top: AW and CE segmentation results for Set$_1$; Bottom: AW segmentation results for Set$_4$.

Figure 4.22: CAg segmentation: box-plot results for the final number of segments generated with different thresholds (X-axis) for Set$_1$;
Figure 4.23: CAg segmentation examples. Results within each column were obtained with the thresholds of $T = 0.1, 0.3, 0.5$. AW Dice/Sensitivity/FP values are listed above corresponding segmentations. Ground truth images are presented in the leftmost column: Cervix boundary outlined in yellow, AW in blue and CE in purple.
III. Analysis of agglomerative clustering with updates

In the agglomerative clustering with updates (AgU) algorithm (Section 3.2.3-II), clusters are merged into single segments whenever one of the clusters reaches the size limit of $N$ members. The following experiment examines the influence of this parameter on the AgU process and compares the results obtained for $N = 3, 4$ and $5$. The NMCut criterion is used for merging. This experiment was performed on $Set_1$.

Figure 4.24(a) presents average Dice and Sensitivity results obtained in the last 50 steps of the AgU algorithm for AW and CE segmentation. Result for $N = 3$ are plotted with a dotted line, results for $N = 4$ are plotted with a dashed line and results for $N = 5$ are plotted with a solid line. The results are similar for $N = 4$ and $5$ and inferior for $N = 3$, for both tissue types. $N = 4$ is hereon used for the AgU algorithm.

The next experiment revisits the issue of graph-cut types, this time with the AgU algorithm. Figure 4.24(b) presents box-plots for the AW and CE segmentation results in the steps of 50, 40, 30, 20, 10 segments within the AgU process. The box-plots compare between the different graph-cuts: WMCut, MinMax-cut and NMCut, presented from left to right, respectively, within each step. AW segmentation results of the NMCut criterion are slightly better (similar) to the results of the MinMax-cut and much better than the results of the WMCut. This observation becomes more significant as the number of segments decrease. The NMCut criterion is considerably better than the MinMax-cut and similar to the WMCut in the case of CE segmentation.

IV. A comparison between the different clustering algorithms

In the last set of experiments we perform a comparison between the different clustering algorithms, Ag, CAg and AgU, within the agglomerative framework. The NMCut criterion is used throughout the analysis. The AgU algorithm is used with a size limit of $N = 4$ components. The step of 50 clusters, that obtained the best results in former experiments, is used for the Ag and AgU algorithms. The CAg algorithm is used with $T = 0.4$, as the number of segments obtained with this threshold possess a median value
Figure 4.24: AW and CE segmentation results via AgU clustering - (a) A comparison across clusters size limits: $N = 3, 4, 5$. Average results for up to 50 segments are shown; (b) A comparison across different graph-cut criterions. Box-plot results for 50, 40, 30, 20, 10 segments are shown. Results for WMCut, MinMax-cut, NMCut, are presented from left to right within each step, respectively ($Set_1$).
of 50 segments (Figures 4.22). The state-of-the-art spectral clustering algorithm with
the normalized-cut graph-cut criterion (Shi and Malik [84]), is used for comparison.
We use the implementation suggested by Ng et al. [67] with the local scaling process
of Zelnik-Manor and Perona [98], with \( K = 50 \) segments. The same \( sKL \) similarity
matrix is used by the different algorithms.

Figure 4.25: A comparison between Spectral clustering, CAg, Ag and AgU (from left
to right within each plot). Presented are box-plots for AW (a) and CE (b) results on
\( Set_1 \) and for AW (c) results on \( Set_4 \).

Figure 4.25 presents box-plot results for AW (a) and CE (b) segmentation quality,
obtained with the different algorithms on \( Set_1 \). The results obtained for AW segmen-
tation quality on \( Set_4 \) are presented in (c). Table 4.8 presents the p-values computed
with a 5% significance level for a paired t-tests that compares the Ag algorithm to the
other clustering algorithms.

The best results in this experiment are obtained by the Ag and AgU algorithms, for
both sets. The results of the AgU algorithm are slightly (not significantly) better than
the results of the Ag algorithm. These two algorithms obtained significantly better
results than the CAg algorithm and the spectral clustering algorithm on \( Set_1 \), which
contains the more complicated AW lesions. \( Set_4 \) results are overall better than \( Set_1 \).

\footnote{1\text{code available at:}http://www.ee.technion.ac.il/~lihi/Demos/SelfTuningClustering.html}
Table 4.8: P-values computed with a 5% significance level for the results presented in Figure 4.25. The p-values are computed for a comparison between the Ag algorithm and the other clustering algorithms: AgU, CAg and Spectral clustering.

The spectral clustering algorithm is next compared to the Ag and AgU algorithms when a different number of segments is used (50, 40, 30, 20, 10 segments). Figure 4.26 presents the results of this comparison. Box-plots for spectral clustering, Ag and AgU results are presented from left to right, respectively, within each group. AW and CE results are presented for Set 1 (top) and AW results are presented for Set 4 (bottom).

The following observations are made: 1) The performance of all three algorithms decrease with the number of segments; 2) The AgU and Ag algorithms outperform the spectral clustering results, in all of the examined cases (this observation is more significant on Set 1 than on Set 4); 3) The AgU results are slightly (but not significantly) better than the Ag results. In the case of AW segmentation with 40 segments, the AgU is significantly better than the other two algorithms.

Figure 4.27 presents example segmentations of the different algorithms in this experiment, with ten segments. Spectral clustering results are presented in column (b), Ag results in column (c) and AgU results in column (d). Manual markings are presented in column (a). AW Dice/Sensitivity/FP results are listed above corresponding results. The cervigrams shown are the same as in the previous experiments.

The following observations are made with regards to examples (I) and (II): 1) Segments generated by spectral clustering are of similar (balanced) size and more convex in shape. The AW region is erroneously merged with the surrounding tissues (e.g. example (II)). A similar behavior was detected for Ag clustering with the $\ell_{\text{MinMax}}$ criterion (Section 4.4.3); 2) The results of the Ag and AgU algorithms with the $\ell_{\text{NMCut}}$ criterion provide elongated segments, which are more balanced in size. These segments
Figure 4.26: Algorithms comparison for a different number of segments (50,40,30,20,10). Box-plots for spectral clustering, Ag and AgU results are presented within each group, from left to right, respectively. Top: AW and CE segmentation results for Set1; Bottom: AW segmentation results for Set4.
capture the non-convex shape of the tissues within the cervix more accurately. The difficult case of Example (III), obtains poor results for all the examined algorithms.

Figure 4.27: Segmentation examples with different clustering algorithms for ten segments. Segments boundaries imposed in black. (a) Ground truth markings; Cervix boundary marked in yellow, AW in blue and Ce in purple; (b) Spectral clustering; (c) Ag; (d) AgU.
4.5 Preliminary tissue classification results

Following the image segmentation phase, each segment can be labeled into one of three possible tissues (AW, SE, CE), using different classification schemes. In this section preliminary classification results are presented. The section includes a comparison between the two frameworks for cervix tissues segmentation: The pixel-based segmentation (Section 3.1) and segmentation via clustering of superpixels (Section 3.2). It provides a summary for the current status of the cervix tissues segmentation task.

Two classification schemes are used in the experiments: In the first scheme, segments and tissues are each represented as a \( d \)-dimensional Gaussian in a selected feature space. Segments are associated with the most similar tissue Gaussian via the KL similarity measure (Equation (3.6)). This scheme is hereon termed the Gaussian-KL (GKL) classifier. Gaussian tissue models are learned from the same training set that was used in previous experiments (Section 4.3).

In the second scheme, a K-nearest neighbor (Knn) classifier is used. In this scheme the segments are represented as single Gaussians (similar to the GKL classifier), but tissues are represented as a collection of Gaussians and not as single ones. Gaussians of the same tissue are all tagged with the same tissue label. In the classification process each segment within the input image is compared with the labeled set of Gaussians via the KL similarity measure. The K-nearest Gaussians are selected and the segment is labeled using a majority voting rule. This scheme is hereon termed the Knn-Gaussian-KL (Knn-GKL) classifier. The collection of Gaussians is learned per tissue out of the different tissue markings within the images of the training set. Each Gaussian is learned from the markings of a different image. The labeled set of Gaussians in our experiments consists of 63 Gaussians (21 Gaussians per tissue).

All the experiments use tissue segmentation to evaluate the classification results. In this scenario, pixels are labeled with the label of their corresponding segment and a segmentation map that includes AW, CE and SE regions, is generated. The segmentation map is next compared to the ground truth expert markings using different
overlap quality measures. Experiments are performed on the images of Set\textsubscript{1} within the automatically detected cervix boundary.

In the first experiment, classification is applied to the hierarchy level of 50 segments generated by the Ag algorithm with the NMCut criterion. The segments in this experiment are represented as Gaussians in $Lb$ feature space. An additional feature combination, $LbR$, is examined, where $R$ is the normalized radial distance between each pixel and the center of the cervix region (computed as the center of mass of the detected cervix boundary). The $R$ feature provides additional information for the classification process, about the relative position of the tissues within the cervix. The experiment compares between the GKL classifier and the Knn-GKL classifier with $K = 1, 5$ and 10 neighbors and examines the influence of the $R$ feature on the classification quality.

Figure 4.28 presents box-plots for the overall Accuracy and the per-tissue Dice segmentation results computed in this experiment. Results in each plot are divided into the $Lb$ and the $LbR$ feature spaces. Box-plot results for the GKL classifier ($G$) and the Knn-GKL classifier with the different $K$ values ($K1, K5, K10$), are shown within each partition. The following observations are made: 1) The addition of the $R$ feature to the $Lb$ feature space improves the results of the Knn-GKL classifier in all of the examined cases. This observation is more significant for the CE tissue, as this tissue is located in the central part of the cervix and for the SE tissue that surrounds it; 2) The best classification results, across all tissue types, are obtained for the Knn-GKL classifier with the $LbR$ feature space and $K = 10$ neighbors. For the CE case the differences are less significant between the different classifiers.

The next experiment compares the classification results obtained with the different image representations used throughout this thesis: pixels, superpixels and segments. Pixels are represented as feature vectors in $LbR$ feature space. Segments and superpixels are represented via Gaussian distributions in the same feature space. The experiment is conducted with the GKL classifier and with the Knn-GKL classifier with $K = 10$ neighbors. For pixel classification the KL similarity within the classifiers is replaced with the labeling procedure of Equation (3.1).
Figure 4.28: Classification results for segments generated via Ag clustering. Box plots of Accuracy and Dice results for the GKL classifier (G) and the Knn-GKL classifier (K1, K5, K10) are presented, with Lb and LbR feature spaces (Set1).

Figure 4.29 presents Accuracy and Dice box plot results for this experiment. The results within each plot are divided into pixels, superpixels and segments, respectively. Box-plots for the GKL classifier (G) and the Knn-GKL classifier (K10), are shown within each partition. The best Accuracy and SE_Dice results are obtained for the Knn-GKL classifier with the segments representation (rightmost column, median SE_Dice = 0.89). The best CE_Dice segmentation results are obtained for the GKL classifier in the segments representation level (median CE_Dice = 0.65). This result indicates that using a single Gaussian in LbR feature space might be adequate for CE tissue modeling. The best AW_Dice results are obtained for the superpixel representation with the Knn-GKL classifier (median AW_Dice = 0.35). This classifier obtained better AW_Dice results in all of the examined representations, as compared to the GKL classifier.

In the last experiment we apply the context-based rules of the pixel-based segmentation framework (Section 3.2) to our classification results and investigate their influence on the classification quality. These rules eliminate small AW and CE regions and smooth out the segmentation maps. They also detect the CE region as a set of
connected components located in the center of the cervix. AW and CE regions eliminated by these rules are replaced with the SE tissue within the segmentation map. This experiment also provides a comparison between the pixel-based segmentation framework and segmentation via clustering of superpixels. For the pixel-based segmentation framework we use both the Lb and the LbR feature spaces. Tissue models in that case are represented as single Gaussians and pixel classification is performed as in the previous experiment. For segmentation via clustering of superpixels we use the hierarchy level of 50 segments generated by the Ag algorithm with the NMCut criterion. Segments are represented as Gaussians in LbR feature space and the Knn-GKL classifier with K = 10 neighbors, is used.

Figure 4.30 presents Accuracy and Dice box-plots results, where each plot is partitioned into results with (right partition) and without (left partition) the context-based rules. Box plots for the pixel-based classification results are marked as pLb and pLbR. Box-plots for the segments classification results are marked as sLbR. Table 4.9 presents
Figure 4.30: Classification results with (right partition) and without (left partition) the addition of context-based rules. Box plots of Accuracy and Dice results are shown for pixel-based classification with the GKL classifier in $Lb$ and $LbR$ feature spaces ($pLb, pLbR$) and for segments classification via the Knn-GKL classifier and the $LbR$ feature space ($sLbR$) ($Set_1$).

The p-values (computed with a 5% significance level) for a paired t-test that compares between different cases within this figure. The following observations can be made: 1) The best results overall are obtained by the segments classification via the Knn-GKL classifier, $sLbR$. The results for this framework are significantly better than the results of the other two methods (median $SE_{Dice} = 0.9$, $CE_{Dice} = 0.65$, $AW_{Dice} = 0.29$). An application of the context-based rules in addition to the $sLbR$ classifier doesn’t improve the results; 2) The context-based rules impose a significant improvement to the pixel-based $CE_{Dice}$ results ($pLb$ and $pLbR$). This result is expected, as the amount of falsely detected CE regions is reduced in that case. Note that the addition of the $R$ feature to the pixel-based scheme is further improved by the context-based rules, which motivates their use; 3) The pixel-based $AW_{Dice}$ results are only slightly improved by the context-based rules, as expected. The addition of the $R$ feature to the $Lb$ feature space slightly improve the results in that case.

Figure 4.31 shows example segmentations obtained in this experiment when the
<table>
<thead>
<tr>
<th></th>
<th>original results - a comparison to sLbR</th>
<th>context-based rules - a comparison to sLbR</th>
<th>original results vs. context-based rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>$sLbR$</td>
<td></td>
<td></td>
<td>0.72</td>
</tr>
<tr>
<td>$pLb$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$pLbR$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$SE_{Dice}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sLbR$</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>$pLb$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$pLbR$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$CE_{Dice}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sLbR$</td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>$pLb$</td>
<td>$\sim 0$</td>
<td>0.16</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$pLbR$</td>
<td>$\sim 0$</td>
<td>0.46</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$AW_{Dice}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sLbR$</td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>$pLb$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
</tr>
<tr>
<td>$pLbR$</td>
<td>0.014</td>
<td>0.047</td>
<td>$\sim 0$</td>
</tr>
</tbody>
</table>

Table 4.9: P-values computed with a 5% significance level for the results presented in Figure 4.30. The p-values are computed for a comparison between the $sLbR$ classification framework and the $pLb$ and $pLbR$ frameworks on the original segmentation results and when the context-based rules are applied. P-values that measure the influence of the context-based rules in the different classification frameworks are shown as well.
context-based rules are applied. Each row includes results for the pixel-based classification framework in $Lb$ ($pLb$) and $LbR$ ($pLbR$) feature spaces and for segments classification via the Knn-GKL classifier in $LbR$ feature space ($sLbR$). The manual markings of the expert are presented on the left image of each example. Accuracy and $AW_{Dice}$ values are listed under corresponding segmentations. We note the following: 1) The maps generated with the segments are smoother as compared to the pixel-based segmentations; 2) CE regions are detected more accurately as compared to AW regions, still some false positives remain; 3) AW regions are over segmented in most of the cases. This result is more significant when the $R$ feature is added to the pixel-based framework; 4) AW regions that remain dark following the illumination correction, might be missed (Example II).

Figure 4.31: Classification examples. Results are shown for the pixel-based classification framework in $Lb$ ($pLb$) and $LbR$ ($pLbR$) feature spaces and for segments classification via the Knn-GKL classifier in $LbR$ feature space ($sLbR$). The context-based rules are applied in all cases. Accuracy/$AW_{Dice}$ values are listed below corresponding segmentations.
Chapter 5

Evaluation of uterine cervix segmentations using ground truth from multiple experts

5.1 Introduction to multi expert evaluation

The NIH cervigram database contains a subset of 939 cervigrams that were each segmented by up to twenty medical experts (Jeronimo et al. [39]). The segmentation was performed using the Boundary Marking Tool software, developed by the National Library of Medicine (NLM) and NCI (Jeronimo et al. [38]). Two clinically important regions were marked by the experts within each image: the cervix boundary and the acetowhite region. Clinical patient information, such as cytology and Human Papillomavirus (HPV) status, was not revealed to the experts. Examples of manual markings, varying per image from one to twenty, can be seen in Figure 5.1. As we consider the segmented images it is evident that several key issues need to be addressed in multiple-expert scenarios: What is the ground truth? Is it the intersection of the markings or their union? Is one expert better than the other? Was the segmentation task a difficult one? As Figure 5.1 illustrates, there are “simple” cases, in which most of the experts agree on the tissue boundaries (Figure 5.1(c),(g)) and more “complex” cases, where the experts have substantially different markings which vary in size and location (Figure
5.1(a),(e)). How can this level of agreement between the experts be quantified? When building an automated system for cervigram segmentation and analysis, how should we quantify the performance of an automated segmentation algorithm as compared to the markings of multiple experts? What are the assessment measures that should be used?

Quantitative evaluation and validation of medical image analysis is a well-known challenge. Several measures for the comparison of automated segmentation results to manual segmentation of a single expert have become a standard in the field (Fenster and Chiu [19], Gerig et al. [21], Udupa et al. [89]). A description of part of these measures is available in Chapter 4. Different works have attempted to handle the above-listed questions for the case of multiple expert data (Chalana and Kim [6], Ladak et al. [50], Martin-Fernandez et al. [63], Warfield et al. [92, 94, 95]). These works focused on generating an average segmentation map using contour-based or area (volume)-based metrics. One of the common methods for example, is to generate the multi-expert ground truth using a simple majority voting rule per pixel [92]. A shortcoming of this method is that it lacks a strategy for determining the number of experts that should agree before the structure (pixel) is considered to be part of the ground truth segmentation. It treats each expert equally without regard to a potential variability in the quality of his segmentation and does not admit use of prior information about the structure being segmented.

A well-known algorithm, that copes with these issues is the STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm (Warfield et al. [94]), which takes a collection of binary segmentations and computes simultaneously a probabilistic estimate of the true segmentation and the performance parameters of each input segmentation. These parameters are computed using the area-based metrics of sensitivity and specificity. STAPLE has been used in the literature in varying application domains, such as generating ground truth maps for Magnetic Resonance Images (MRI) of the brain (Cuadra et al. [11], Rohlfing et al. [79]), 3D medical structures (Cates et al. [5]) and open curves of vascular structures (Jomier et al. [42]). It has been used
for constructing a brain MRI atlas for two-year-old children (Josh et al. [43]), and in combining two-class maps to obtain a complete segmentation of a brain tissue (Li et al. [56]). It has also been used for object recognition (Mattern et al. [65]).

No work has analyzed multi-expert segmentations of uterine cervix images (specifically via STAPLE). Furthermore, the issues of segmentation complexity and how it is measured within the STAPLE framework were not addressed. The objectives of the current work are to use the STAPLE algorithm in order to combine different expert markings and to generate a single ground truth map for cervigrams. It presents a thorough analysis of uterine cervix segmentation via STAPLE, utilizing different analysis schemes. A new segmentation-complexity measure is defined based on the multi-expert ground truth map. The performance parameters of sensitivity and specificity are the common assessment measures used to evaluate segmentation quality in the STAPLE literature (some additional methods are presented and discussed by Zou et al. [101]). These measures are known to possess incommensurate magnitudes, as they represent percentages from different populations of pixels (the object and the background) which are very different in size (Cates et al. [5]). The current work addresses this difficulty and demonstrates its effect on the STAPLE performance. It then defines an accuracy measure to evaluate the results of automatic segmentation algorithms as compared to the multi-expert ground truth map. The accuracy measure is used for the evaluation of two automated algorithms for cervix boundary detection. The focus of the work is on the cervigram database, but the methods proposed are general.

The rest of this chapter is organized as follows: The STAPLE algorithm is described in Section 5.2. Its sensitivity to the size of different populations is discussed. Methods for performance analysis based on the STAPLE output are presented in Section 5.3. Experimental results on the cervigram database are described in Section 5.4. A discussion concludes the chapter in Section 5.5.
5.2 The STAPLE algorithm

The STAPLE algorithm [94] takes a collection of binary image segmentations as an input (a single segmentation per expert). The object pixels within these segmentations are marked as one and the background pixels as zero. The algorithm simultaneously computes: (1) a probabilistic estimate of the true segmentation and (2) a measure of the performance level represented by each input segmentation. The algorithm is formulated as an instance of the expectation-maximization (EM) algorithm (Dempster et al. [12]). The performance levels, or quality achieved by each expert, are represented by the sensitivity and specificity parameters. The sensitivity ($p_j$) of expert $j$ represents the “true positive fraction”: $p_j = Pr(D_{ij} = 1|T_i = 1)$. The specificity ($q_j$) of expert $j$ represents the “true negative fraction”: $q_j = Pr(D_{ij} = 0|T_i = 0)$, where $D_{ij}$ is the decision made by expert $j$ for pixel $i$ (1 meaning: present in the expert’s segmentation and 0, absent) and $T_i$ is the hidden value of the true segmentation at pixel $i$.

The EM algorithm estimates the performance level parameters ($p$, $q$) while maximizing the complete data log likelihood function. It iterates as follows: In the E-step
the unobserved true segmentation is computed as:

\[
f(T_i | D_i, p^{(k-1)}, q^{(k-1)}) = \frac{\prod_j f(D_{ij} | T_i, p_j^{(k-1)}, q_j^{(k-1)}) f(T_i)}{\sum_{T'_i} \prod_j f(D_{ij} | T_i, p_j^{(k-1)}, q_j^{(k-1)}) f(T'_i)},
\]

where \(f(T_i)\) is the prior probability for tissue \(i\) and \(k\) is the iteration step. Considering a binary segmentation, factoring over all the experts and using the definitions for \(p_j\) and \(q_j\), the following formulas are defined:

\[
a_i^{(k)} \equiv f(T_i = 1) \prod_j f(D_{ij} | T_i = 1, p_j^{(k)}, q_j^{(k)}) = f(T_i = 1) \prod_{j: D_{ij} = 1} p_j^{(k)} \prod_{j: D_{ij} = 0} (1 - p_j)^{(k)},
\]

\[
b_i^{(k)} \equiv f(T_i = 0) \prod_j f(D_{ij} | T_i = 0, p_j^{(k)}, q_j^{(k)}) = f(T_i = 0) \prod_{j: D_{ij} = 0} q_j^{(k)} \prod_{j: D_{ij} = 1} (1 - q_j)^{(k)},
\]

where \(j : D_{ij} = 1\) denotes the set of indices for which the decision of the rater at pixel \(i\) has the value 1. Using these formulas, a compact expression for the unobserved true segmentation at each pixel, \(W_i\), is defined:

\[
W_i^{(k-1)} \equiv f(T_i = 1 | D_i, p^{(k-1)}, q^{(k-1)}) = \frac{a_i^{(k-1)}}{a_i^{(k-1)} + b_i^{(k-1)}}.
\]

The experts performance level parameters are estimated in the M-step using the following equations:

\[
p_j^{(k)} = \frac{\sum_{i: D_{ij} = 1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}}, \quad q_j^{(k)} = \frac{\sum_{i: D_{ij} = 0} (1 - W_i^{(k-1)})}{\sum_i (1 - W_i^{(k-1)})}.
\]

The sensitivity estimator, \(p_j\), can be interpreted as the ratio of the \(j\)th expert true positive detections to the total amount of the structure \(T_i = 1\), where in both cases each pixel is weighted by \(W_i\): the strength of belief in \(T_i = 1\). Similarly, the specificity estimator, \(q_j\), can be interpreted as an estimator for the specificity given a degree of belief in the underlying \(T_i = 0\) state.
The unobserved true segmentation computed in the E-step is a probability map where each pixel is assigned the probability of being part of the segmented object according to (1) the amount of agreement among the experts and (2) the performance levels of the experts. This map is regarded as the “multi-expert ground truth segmentation” generated by STAPLE. Figure 5.1 shows examples of multi-expert ground truth maps that correspond to the expert markings for both the acetowhite region (b, d) and the cervix boundary (f, h). Pixel probabilities are color-coded from blue (low probability - zero) to red (high probability - one). The intersection of all experts’ markings is colored red (with the highest probability value) as expected.

In the current task of cervigram segmentation, the object area (the cervigram region) is relatively small as compared to the area of the background. The amount of pixels for which the experts disagree when marking the object, is even smaller as compared to the background. In such cases the dynamic range of the specificity and the sensitivity values is incommensurate. The specificity values obtain higher values with a narrower dynamic range, as they are computed with respect to the background pixels (Equation (5.5)). This behavior has a major influence on the estimated ground truth segmentation: The $b_i$ values (Equation (5.3)) are small compared to the $a_i$ values and the resulting $W_i$ (Equation (5.4)) values are higher than expected. An exact similar dynamic range for the two performance measures is obtained only when the amount of pixels with zero probability to be part of the object is equal to the amount of pixels with probability one to be part of the object. Such a case seldom happens in real-life segmentations.

We use the following procedure to obtain more comparable performance measures: The union of the different expert markings is considered to be the object area. A similar number of pixels is drawn out of the background to define the new background area. The idea of redefining the background area prior to the manual or automatic segmentation process was previously suggested in order to improve the segmentation results and the STAPLE-based validation (Zou et al. [101]). The positive influence of this modification on the STAPLE’s ground-truth output, is demonstrated here.
Figure 5.2 illustrates the suggested background modification and its influence on the STAPLE algorithm output. Figure 5.2(a) presents five overlapping segmentation masks of ellipses, each with a different orientation and size, the rest of the image is considered to be the original background area.

Figure 5.2(d) presents the new scene generated by the suggested background modification. The union of the masks is colored red. The pixels drawn out of the background, colored green, are equally distributed around this region and are of the same amount as the red region. The rest of the pixels, colored blue, are masked out and ignored throughout the rest of the computations. Figures 5.2(b) and 5.2(c) are the ground truth segmentation maps generated by the STAPLE algorithm when using the original background area and the modified scene, respectively. Figures 5.2(e) and 5.2(f) are corresponding histograms, reflecting the distribution of the probability values within.
Table 5.1: STAPLE simulation: sensitivity and specificity values computed for each of the segmentation masks of Figure 5.2 using the original image size \((p, q)\) and the modified scene \((p_m, q_m)\).

<table>
<thead>
<tr>
<th></th>
<th>(I_1)</th>
<th>(I_2)</th>
<th>(I_3)</th>
<th>(I_4)</th>
<th>(I_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>1.00</td>
<td>0.48</td>
<td>0.44</td>
<td>0.31</td>
<td>0.24</td>
</tr>
<tr>
<td>(q)</td>
<td>0.94</td>
<td>0.99</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(p_m)</td>
<td>1.00</td>
<td>0.67</td>
<td>0.26</td>
<td>0.46</td>
<td>0.36</td>
</tr>
<tr>
<td>(q_m)</td>
<td>0.68</td>
<td>0.97</td>
<td>0.86</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.1 presents the sensitivity and specificity values computed for each of the masks when using the original image \((p, q)\) and the modified scene \((p_m, q_m)\). Without the scene modification the specificity values are higher and with a narrower dynamic range. Note the significant change in the specificity values of expert 1, with the lowest specificity. The sensitivity values also do not accurately reflect the level of agreement between the experts. For example the sensitivity of expert 3 is higher than those of experts 4 and 5. When the modified scene is used, the range of the specificity values is wider and expert 3 attains the lowest sensitivity.
5.3 Performance analysis based on multi-expert ground truth maps

The multi-expert ground truth (GT) map generated by STAPLE is a fuzzy probability map that includes the probability per pixel to belong to the object region within the image. In this section we propose to use the map for quantifying segmentation complexity (Section 5.3.1) and for quantitatively evaluating the performance of an automated segmentation algorithm (Section 5.3.2).

5.3.1 Measuring segmentation complexity

We propose to compute a set of measures, or “descriptors”, to represent the complexity of a given segmentation task. The underlying assumption is that the variability of pixel probabilities within the GT map indicates the level of disagreement between the human experts. This, in turn, is an indication for the increased complexity of the segmentation task. The probability values within the GT map are linearly stretched prior to the descriptors computation to enable a comparison of their distributions across the images. The highest probability value is set to one and lowest value to zero. The probability value of a pixel \( i \) within the GT map, \( I \), is denoted by \( W_i \). The descriptors are computed only for the object area, as defined by the STAPLE output (\( W_i > 0 \)).

We propose the following descriptors:

1. **Entropy:**

   Entropy is a well known measure of distribution homogeneity. It is computed here using the histogram representation of the distribution within the GT map:

   \[
   \text{entropy}(I) = - \sum_{i=1}^{N} h_i \ast \log(h_i)
   \]

   where \( h_i \) is the probability of bin \( i \) of the histogram (\( \sum h_i = 1 \)) and \( N = 100 \).

   An intuitive understanding of entropy relates to the amount of uncertainty in the segmentation: A GT map that contains a single probability value, as is the case
in complete overlap of the expert segmentations, has an entropy value of zero (no uncertainty). Disagreement between the experts generates additional probability values within the GT map, which leads to a broader distribution and to a higher entropy value.

2. **Standard deviation (STD):**

The width of a distribution can be measured using its standard deviation:

\[
STD(I) = \left( \frac{1}{n-1} \sum_{i=1}^{n} (W_i - W_m)^2 \right)^{\frac{1}{2}},
\]

where \( n \) is the number of pixels considered to be an object and \( W_m \) is the mean of their probability values. A low \( STD \) value is associated with a narrow distribution of probabilities within the GT map and vice versa.

3. **Entropy or Standard Deviation scaled by Mean:**

While the entropy and the standard deviation well represent the spread of the distribution, they do not represent the probability values themselves. This information is important when measuring segmentation complexity. A narrow distribution may be located in the high-probability range, or in a low-probability range. The first case corresponds to GT maps with large areas of strong agreement and the second case corresponds to GT maps with large areas of strong disagreement between the experts. (Examples and discussion will be provided in Section 5.4). In order to cope with such cases, we propose a *normalized* set of descriptors. In this set the entropy and the standard deviation are scaled by the square of the distribution’s mean. These descriptors are termed “entropy scaled by mean” \((ESM)\) and “standard deviation scaled by mean” \((SSM)\), respectively and are defined as:

\[
ESM(I) = \frac{\text{entropy}(I)}{\text{mean}(I)^2}; \quad SSM(I) = \frac{\text{std}(I)}{\text{mean}(I)^2}.
\]

We now have four types of descriptors. We classify a given segmentation task as
“simple” or “complex”, relative to a selected set of these descriptors. We use either a thresholding or a clustering approach to carry out this classification. A threshold for the complexity can be learned from a training set of segmentations, following which each new segmentation can be categorized using the selected threshold. We use this method for the ESM (SSM) descriptor. For the clustering scheme, we use a 2-D feature space of the entropy and the mean descriptors. Training data is used to cluster the space into varying complexity levels. Based on the 2-D clustering of the complexity feature space, each new image input to the system can be categorized as less or more complex (depending on its own GT map descriptor set). It is also possible to analyze the variability of the segmentation complexity across the images within the database. In addition, the complexity of segmenting different regions within the cervix can be compared, thus distinguishing between easy and difficult segmentation tasks.

5.3.2 Evaluating automatic segmentation results

Given a new segmentation map, created independently of STAPLE, it may be desirable to compare it quantitatively to the STAPLE multi-expert ground-truth. Such an analysis can be used to assess the performance of an automated segmentation algorithm and to compare the results of different algorithms. This analysis can be made using the following methods:

1. Computation of the sensitivity and specificity performance levels of the new segmentation as compared to the multi-expert ground truth, using Equation (5.5) (Warfield et al. [94]).

2. Computation of the accuracy (Fawcett [18]) of the new segmentation as compared to the multi-expert ground truth. The accuracy of a given classifier, defined on a binary set of samples with positive $P$ and negative $N$ labels is computed as the total correct fraction: $\frac{TP + TN}{P + N}$, where $TP$ and $TN$ are the amount of true positives and true negatives detected by the classifier. In the current case, the multi-expert ground truth is a set of real numbers, $W$, which define the probability for a
positive label; denote the new segmentation by $D$. Then we compute \textit{accuracy} as:

\[
\text{accuracy} = \frac{\sum_{i: D_i = 1} W_i + \sum_{i: D_i = 0} (1 - W_i)}{N},
\]

where $D$ and $W$ are treated as vectors and $N$ is the number of samples being considered. Higher \textit{accuracy} values indicate better correspondence to the ground truth (Cates \textit{et al.} [5]).

In the first method, the sensitivity and specificity parameters computed for a specific image are compared to the parameters attained by the human experts for that image. Each of the parameters is evaluated separately. This method has two main drawbacks: First, as the number of experts increases, it is more complicated to rank their results and compare them to the results of the algorithm. A single measure is more appropriate in that case. Second, as demonstrated in Section 5.2, the dynamic range of the sensitivity and specificity values depends on the relative size of the object and background within the image. Thus, care should be taken when combining them into a single measure. In addition, these measures can be used to compare results within a single image (of the same data [5]), but not across the images in the database, as the size of the objects varies considerably. A statistical evaluation of segmentation algorithm results is applicable only when the comparison is performed between different algorithms and on the same data.

In an earlier work (Lotenberg \textit{et al.} [62]), the \textit{F-measure} (Rijsbergen [90]) was suggested in order to combine the specificity $q$ and sensitivity $p$, into a single value. The \textit{F-measure} is defined as the weighted harmonic mean of the two parameters:

\[
F = \frac{pq}{\alpha p + (1 - \alpha)q}, \quad \alpha = 0.5.
\]

Being dependent on sensitivity and specificity, the \textit{F-measure} is strongly affected by the relative object to background area. This measure also assumes a similar dynamic range for sensitivity and specificity within a single image (by setting parameter $\alpha$ to 0.5), which is seldom true in the current case. Other measures that combine specificity and
sensitivity into a single measure suffer from the same faults. These measures include the shortest distance from the \((0, 1)\) corner, used in Receiver Operating Characteristic (ROC) analysis (Fawcett [18]) and the mean predictive value \((PV)\) (Warfield et al. [94]) that reduces to \(\frac{p+q}{2}\) in the binary case.

The \textit{accuracy}, used in the current work accounts for the amount of accurately detected labels, as compared to the image size and not the size of the different regions within it. The modification of the background area, in order to balance the object/background proportions, makes this measure even less sensitive to their relative size, as compared to the other options.

\[
\begin{align*}
F & \quad 0.97 & 0.65 & 0.60 & 0.47 & 0.39 & I_1, I_2, I_3, I_4, I_5 \\
F_m & \quad 0.81 & 0.79 & 0.40 & 0.63 & 0.53 & I_1, I_2, I_3, I_4, I_5 \\
PV & \quad 0.97 & 0.74 & 0.71 & 0.66 & 0.62 & I_1, I_2, I_3, I_4, I_5 \\
PV_m & \quad 0.84 & 0.82 & 0.56 & 0.73 & 0.68 & I_1, I_2, I_3, I_4, I_5 \\
acc & \quad 0.945 & 0.959 & 0.947 & 0.954 & 0.950 & I_2, I_4, I_5, I_3, I_1 \\
acc_m & \quad 0.73 & 0.92 & 0.77 & 0.91 & 0.90 & I_2, I_4, I_5, I_3, I_1
\end{align*}
\]

Table 5.2: STAPLE simulation. \textit{F-measure}, mean predictive value and \textit{accuracy} results computed for each of the segmentation masks of Figure 5.2 for the original image size \((F, PV, acc)\) and the modified scene \((F_m, PV_m, acc_m)\). Corresponding ranking order of segmentations, from most to least similar, is included (left to right).

Table 5.2 presents the \textit{F-measure} \((F)\), mean predictive value \((PV)\) and the \textit{accuracy} \((acc)\) results computed for the different segmentations, \(I_1, \ldots, I_5\), within the simulation of Figure 5.2. These results illustrate the benefits of using the \textit{accuracy} measure. The values are computed using the original image size and the modified scene. A higher value indicates a more accurate segmentation, as compared to the STAPLE-generated, multi-expert ground truth. In order to identify the segmentation that is most similar to the multi expert ground truth, the different values are sorted in decreasing order and the different segmentations are ranked accordingly. The desired ranking according to our perceptual understanding may be: \(I_2, I_4, I_5, I_3, I_1\), where \(I_2\) is the most similar segmentation to the ground truth.
The following observations can be made: 1) The *accuracy* measure obtains a similar ranking of segmentations for the full image area \((acc)\) and for the modified scene \((acc_m)\). This observation indicates reduced sensitivity of the *accuracy* measure to the relative size of the object and the background; 2) The ranking of the *accuracy* measure appears to correspond with our perceptual understanding; and 3) The dynamic range of the *accuracy* is increased when using the modified scene. This generates a better distinction between the quality of the different segmentations within a single image and provides additional support for the scene modification suggested in Section 5.2.

The above observations are not exhibited in the case of the *F*-measure and the *PV*. The change in the dynamic range of the sensitivity and specificity values in the modified scene, as compared to the full image, affects their corresponding descriptor \((F, PV)\) values and ranking. This indicates that these measures are more sensitive to the relative proportions of the object and background. An additional observation relates to the ranking order itself, where \(I_1\) is wrongly ranked as most similar to the GT. This misplacement occurs because a similar range of sensitivity and specificity is assumed in the computation. The range of the specificity values, improved by the modified scene (Table 5.1), is still narrower than that of the sensitivity.

### 5.4 Experiments and Results

A set of experiments was conducted in order to evaluate the proposed computational measures and analysis schemes. The database used contains a set of 932 manually segmented cervigrams out of the 939 cervigrams of the NCI database (Jeronimo *et al.* [39]). Seven cervigrams were discarded since their segmentations exhibited no overlap whatsoever between the human experts. The NCI database was divided into two main groups. One group contains 20 cervigrams that were marked by twenty medical experts. The second group contains the remaining 912 cervigrams, with each marked by two experts out of the twenty medical experts. The markings of two regions are examined: the acetowhite region and the cervix boundaries. A version of the STAPLE algorithm
CHAPTER 5. EVALUATION OF MULTIPLE-EXPERTS MARKINGS

that considers the modified scene (Section 5.2) was used throughout the analysis to generate the multi-expert ground truth\(^1\).

5.4.1 Evaluation of the segmentation complexity descriptors

We start by examining the correlation between the proposed segmentation complexity descriptors (Section 5.3.1) and the level of agreement among the experts, as given by the multi-expert ground truth generated by STAPLE. Figure 5.3 shows examples of the multi-expert ground truth segmentation for both the cervix boundary and the acetowhite region. Histograms of the probability values within each map are presented under corresponding examples (for the object region only). The segmentation complexity descriptors computed for each of these examples are listed. The following observations can be made:

- The **Entropy** measures the distribution of the probability values within the map, without taking into account their magnitudes. This results in cases such as the one presented in Figure 5.3(g), where the entropy is very low but the disagreement between the two experts is clearly visible.

- The **Standard deviation (STD)** has a similar deficiency: it accounts for the distribution of the probability values around their mean, but does not consider the mean value itself. Like the entropy, the standard deviation will fail (i.e., take on low values, even though the disagreement between the experts is high) in cases such as the one presented in Figure 5.3(g).

- The **Mean** descriptor is the average value of the probabilities within the ground truth map. A correlation can be detected between high mean values and strong agreement between the experts. The mean value, however, lacks the ability to differentiate between cases with similar mean and different distributions.

- The **Entropy Scaled by Mean (ESM)** combines the benefits of both the en-

\(^1\)The original STAPLE algorithm is available via the ITK toolkit (http://www.itk.org/)
tropy and the mean descriptors and successfully differentiates among the different levels of agreement in all of the presented examples. Low values of ESM are correlated with high levels of agreement. The **STD Scaled by Mean (SSM)** attains similar results.

From these observations it is evident that a high level of agreement between the experts is captured well by the ESM or the SSM descriptors, where both the distribution and the mean of the probability values within the ground-truth segmentation are considered. It is important to note that the ESM/SSM values are strongly influenced by the number of expert markings, since a larger number of experts may produce a wider range of probability values within the multi-expert ground truth. A more reliable comparison would be between images that were marked by the same number of experts.

In a second experiment, we evaluate segmentation complexity by examining clusters in the two-dimensional feature space of entropy and mean. In this feature space, low entropy and high mean values are correlated with easier cases, where expert agreement is high. Figure 5.4 presents a scatter plot of the entropy and the mean descriptors computed for 100 expert segmentations. These segmentations were randomly selected out of the 912 cervigrams that were marked by two experts. The experiment was conducted for segmentation of (a) the cervix boundary and (b) the acetowhite region.

The distribution of the cervix boundary segmentations, Figure 5.4(a), is mainly concentrated within the low-entropy-high-mean region of the feature space. This indicates a strong agreement among the experts within most of the cervigrams. In the distribution of the acetowhite segmentations, Figure 5.4(b), three main groups can be detected: Group A includes the low-entropy-low-mean segmentations. Group B includes the low-entropy-high-mean segmentations and Group C includes the high-entropy-mid-mean segmentations. This may be interpreted as follows: Group A corresponds to segmentations with strong disagreement among the experts. Group B corresponds to segmentations with strong agreement, and Group C corresponds to segmentations with an intermediate level of disagreement among the experts. Figure 5.5
Figure 5.3: Examples for multi-expert ground truth data for the cervix boundary (top row) and the acetowhite region (bottom row). (a),(b),(e),(f): examples of agreement among experts; (c),(d),(g),(h): examples of disagreement among experts; Corresponding histograms and complexity descriptors are presented under each example.
shows ground-truth segmentation examples for each of these groups along with their ESM values. From top to bottom, each row shows examples for groups A, B and C, respectively. The distinction between the maps in the entropy-mean feature space is highly correlated with the level of agreement between the experts, visually detected in the maps within each group (where the red color corresponds to regions of strong agreement). The ESM descriptor attains the lowest values within the images of group B, where the level of agreement between the experts is high, as expected. The distinction between groups A and C is less evident when using the ESM descriptor.

The images within the three groups were presented to a medical expert who was asked to describe the visual appearance of the acetowhite regions within them. According to the expert most of the acetowhite regions within group A are not visibly clear (described as pale, with diffused and weak acetowhitening). This may explain the poor level of agreement between the experts detected in these images. The distinction between the images within groups B and C is less evident, but most of the acetowhite regions within these groups are described to be very clear and well delimited.
Figure 5.5: Example segmentations for the three groups observed in the entropy-mean feature space of the acetowhite region (Figure 5.4(b)). Top row: group A - low-entropy-low-mean; Middle row: group B - low-entropy-high-mean; Bottom row: group C - high-entropy-mid-mean. ESM values are listed below corresponding results.
CHAPTER 5. EVALUATION OF MULTIPLE-EXPERTS MARKINGS

5.4.2 A comparison between the complexity of the acetowhite and cervix boundary segmentation tasks

The complexity descriptors can be used to assess the segmentation complexity of different images as well as of different segmentation tasks. In the following experiment, the complexity of segmenting the acetowhite region is compared to the complexity of segmenting the cervix boundary. Figure 5.6 presents the distributions of the ESM descriptor for the acetowhite segmentation (a), and for the cervix boundary segmentation (b). The distributions were computed for images that were marked by two experts. The cervix boundary segmentation has a narrower distribution with a lower mean value. This indicates strong agreement among the experts in most of the cases and suggests that the cervix boundary segmentation task is the easier task. Figures 5.6(c) and (d) show scatter plots in entropy-mean feature space for images that were segmented by more than ten observers. The cervix boundary segmentation results, (d), are concentrated in the low-entropy-high-mean region. This reflects the strong agreement between the experts in all cases. The scatter of the acetowhite segmentation results in (c) is more spread out, thus indicating a larger disagreement across the different cases and, correspondingly, a more complex segmentation task. Similar results can be detected in the scatters presented in Figure 5.4, in which the results of 100 cervigrams are presented.

5.4.3 Evaluation of automatic cervix boundary segmentation

We compared two algorithms for cervix boundary detection, using the accuracy measure defined in Section 5.3.2. The first algorithm (algorithm I) detects an initial coarse region of interest (ROI) located around the cervix region (Section 2.1.1). The second algorithm (algorithm II) is based on the results of the active contour framework for cervix boundary detection presented in Section 2.2.1, where only the data-term is activated. Algorithm I is used to initialize algorithm II and is therefore expected to have inferior results.
Figure 5.6: Top row: Distribution of the ESM value for images that were marked by two experts: (a) acetowhite region, mean = 1.93 (602 images); (b) cervix boundary, mean = 0.93 (636 images). Bottom row: Scatters of segmentation complexity in the entropy-mean feature space, for images with more than ten experts markings: (c) acetowhite region (16 images); (d) cervix boundary (20 images).
A comparison was conducted among the following assessment measures: sensitivity (p), specificity (q) and accuracy. All measures were computed using the multi-expert ground truth generated by STAPLE. Only images where the markings of two experts are available were used. Table 5.3 presents the mean and the standard deviation values attained for each of these measures for the two segmentation algorithms tested. Algorithm I attains very high sensitivity values but they are correlated with very low specificity values. These results indicate that the cervix region is always located within the detected ROI but that large portions of the background are also included. Algorithm II, which was designed with the goal of obtaining a more accurate delineation of the cervix boundary, significantly reduces the amount of falsely detected regions. This comes at the expense of missing some of the cervix region pixels. The accuracy measure, which combines the detection quality of the cervix region and the background into a single measure, favors algorithm II as expected.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Algorithm I (mean, std)</th>
<th>Algorithm II (mean, std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.98 (0.03)</td>
<td>0.87 (0.1)</td>
</tr>
<tr>
<td>q</td>
<td>0.38 (0.22)</td>
<td>0.75 (0.19)</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.65 (0.12)</td>
<td>0.8 (0.09)</td>
</tr>
</tbody>
</table>

Table 5.3: Evaluation of two algorithms for cervix boundary detection using different performance measures. Mean and standard deviation results for 636 images are presented (mean(std)).

Figure 5.7 demonstrates the benefits of the accuracy over the F-measure and the mean predictive value (PV), when more than two expert segmentations are available. In this example the three quality measures are used to assess the performance of the experts themselves, as compared to the STAPLE-generated, multi-expert ground truth. The original cervigram with the different markings is presented in (a), and the ground truth segmentation generated by STAPLE is presented in (b). The F-measure, PV and accuracy, computed for each of the available expert segmentations, are listed on top of corresponding maps in (c),(d) and (e), respectively. The maps are sorted in decreasing order of similarity to the STAPLE ground truth, beginning with the most similar map. The sorted maps in (c) and (d) demonstrate the tendency of the F-measure and the
PV to favor larger segmentations. Note the second and third most similar maps in (c): These maps correspond to regions that were marked by single experts and are certainly not within the region of highest probability in the GT map. The sensitivity in these cases is high, as the maps include most of the ground truth region. The specificity is not low enough when compared to the other cases, due to its narrow dynamic range. Both the \textit{F-measure} and the PV are strongly affected by the sensitivity, which leads to erroneous results. The sorted maps in (e), where the \textit{accuracy} measure is used, have a ranking closer to our intuition. The most similar maps in this case correspond to the region of high probability within the GT map, where the level of agreement between the experts is high.

5.5 Discussion

This part of the thesis focuses on generating reliable multi-expert ground truth for the cervigram segmentation task. In addition several descriptors based on the output of the STAPLE algorithm are discussed, including the ESM and SSM descriptors that measure segmentation complexity of a single image, and the \textit{accuracy} measure that evaluates the performance of automated segmentation algorithms as compared to the markings of multiple experts.

The main results of this work are:

- The ESM and SSM descriptors are correlated with the levels of agreement among experts. We have shown the superiority of these descriptors over alternatives such as simple entropy, standard deviation and mean probability measures. The evaluation of segmentation complexity in the entropy-mean feature space was shown to be more accurate than the ESM descriptor when trying to distinguish between different types of disagreement among the experts.

- The ability of the \textit{accuracy} measure to evaluate the results of automated segmentation algorithms was demonstrated and this measure was shown to provide
Figure 5.7: (a) Experts markings imposed on original cervigram. (b) Multi-expert ground truth generated by STAPLE. (c),(d),(e) Sorted segmentation maps according to F-measure, PV and accuracy, respectively. The maps are sorted in decreasing order from most to least similar to the ground truth segmentation.
a reliable evaluation that factors in the detection quality of the object, as well as that of the background, without being too sensitive to their relative sizes.

- The ESM and the entropy-mean feature space were used to characterize the complexity of segmenting acetowhite lesions versus segmenting cervix boundaries. In all of the presented experiments, the acetowhite segmentation was shown to be a more complex segmentation task, with a larger amount of disagreement among experts. This result can be explained by the fact that the acetowhite tissue may consist of multiple regions distributed across the cervix, the tissue is visually more difficult to detect, and it has less well-defined boundaries. The cervix region, on the other hand, is a single connected region that is clearly visible within the cervigram.

The task of automatic uterine cervix image analysis is in its preliminary stages. Detection and segmentation of cervigram tissues is very challenging due to the large diversity of the cervigram images within the database and the different artifacts present in the cervigrams. Tuning algorithms to the segmentation characteristics of a single expert would be unsatisfactory, due to the large multi-expert variability that exists. The complexity definition that we have proposed can be used to classify a database into “simple” and “complex” images. This may aid in the performance evaluation and analysis (per complexity group) of automated segmentation algorithms being developed (as presented in Section 4.4).

We also conducted an initial qualitative comparison between the visual appearance of the acetowhite lesions (described by a single expert) and the segmentation complexity specified by the clusters in the entropy-mean feature space. In this comparison a correlation was detected between lesions that are difficult to detect and images that are complex to segment (where the agreement between the experts was poor). In future work we plan a more thorough analysis of the correlation between segmentation complexity and other medical findings available in the NCI database.

Finally, when ground truth markings from multiple experts are available, the com-
plexity of uterus cervix image segmentation can be evaluated using measures such as the ESM or the two-dimensional entropy-mean feature space, presented in this work. Automatic segmentation algorithms can be analyzed with respect to the multi-expert ground truth map using the specified accuracy measure. The focus of this work are the cervigram images, but the methods presented are general and can be applied to a variety of medical image archives and application domains.
Chapter 6

Discussion and conclusions

This research presents an automated analysis framework for uterine cervix images. The proposed scheme is hierarchical: it first defines a rough estimate for the cervix region within the image and focuses the analysis within this region. In the second processing step, the specular reflection artifacts are detected and eliminated. A more exact delineation of the cervix boundary is then computed, followed by an illumination correction and intensity normalization step. The feasibility of two segmentation frameworks is examined next for tissues segmentation within the cervix region.

Each step in the presented framework is a result of an in-depth study and investigation of the cervigrams and has its own significance. It is treated separately, using appropriate features and specific segmentation algorithms, with the goal of achieving robustness in the segmentation. A specific summary that includes conclusions for each step, discusses the results and presents future directions, is provided in Section 6.1. General conclusions for the research and future directions are presented in Section 6.2.
6.1 Discussion and conclusions for the automated cervix analysis framework

I. Cervix boundary detection

This research presents a novel approach for cervix boundary detection, based on geometrical curvature characteristics of the image and segmentation via an active contour framework. Two methods are presented for incorporating prior shape information into this framework: the first method is based on shape features distribution using a circular prior. The second method is a novel method for embedding an elliptical prior, which uses an implicit shape representation via a level-set function. The presented approach is shown to outperform an existing clustering-based method in the task of cervix boundary detection. The incorporation of prior shape information is shown to further improve the boundary detection quality.

The results obtained with the two shape priors on the entire test-set are similar, as there are other factors that influence the curve evolution process. One of these factors is the initial, data-driven curve and its position relative to the true cervix boundary. This position varies considerably across the cervigram images due to their large content variability. The position of the initial curve has a strong influence on the active contour process, as the process tends to converge into local optimum. A more focused experiment that isolates the effect of the shape-prior term on the segmentation quality, indicates the superiority of the elliptical-prior method over the circular-prior method. The elliptical-prior method (using the level set framework) is also shown to have a reduced order of complexity, $O(x)$, as compared to the complexity, $O(x^2)$, of the circular-prior method (and the corresponding curve-based functional).

The detection of an accurate cervix boundary is critical for the steps of landmark extraction and tissue segmentation, as the regions outside the cervix provide misleading tissues information and landmarks within the cervix (such as the os, Zimmerman et al. [99], Greenspan et al. [29]) are often identified using their relative position from the boundary. The segmentation results obtained in this work were found satisfactory
when compared to the manual expert markings, thus constituting a good basis for further content analysis. Consistency in the results was shown across different image test sets.

An open issue that should be considered in future research is the matching of the shape model to the problem at hand. Although the manual ground-truth markings have a clear tenancy to resemble a circular or elliptical shape, this is not the case for all images. In part of the images, the segmentation quality when using the presented shape models will suffer from lack of resemblance to the ground-truth. There are also cases where one prior overestimates the cervix region while the other prior underestimates it. Future work should include the combination of the two methods in order to produce better results. In addition, different ways for per-image parameter tuning should be investigated, as a single set of parameters for the active contour framework might be too restrictive for the current large and diverse database. Parameter tuning per image may further improve the results.

II. Illumination correction and intensity normalization

The illumination correction and intensity normalization process presented in this work introduces the following key contributions to the cervix analysis framework: 1) The illumination correction step enables a substantial augmentation of the segmentation performance on a per-image basis. The amount of falsely detected AW regions is considerably reduced following this step; 2) The intensity normalization step enables the shift from per-image unsupervised modeling of cervix tissues to supervised modeling of tissues via a training-set of images. These models are used for pixel-based tissues segmentation and for classification of superpixels and segments, in successive steps.

The illumination correction quality using the presented GEM algorithm was shown to provide better results than the Retinex algorithm. This may be explained by the fact that the GEM-based illumination correction uses prior knowledge regarding the image content: Cervigram images are known to contain a number of relatively smooth regions and the illumination artifact can be correlated with a large variance within
The GEM algorithm was originally developed for bias correction in MR brain images, for which many additional methods exist. The choice of a specific method in that case is based on the problem at hand, the source and the quality of the data (Hou [35]). In the current work we have chosen to use the GEM method and have found that it supports an augmentation in segmentation results. A comparison between GEM and other methods for illumination correction of the cervigram data might be adequate as future work.

III. Probabilistic pixel-based cervix tissue segmentation

The research presents a probabilistic pixel-based segmentation framework for cervix tissues, via mixture of Gaussians modeling. An extensive set of experiments tested a wide range of tissue models: From a totally unsupervised model based on the observed image samples, to a fully supervised model learned from a training set, with some focal points in-between that weight both ends during the model learning phase. In these experiments, the SE and the CE segmentations were better when both the observed image samples and the prior were considered in similar proportions. The overall accuracy results attained a similar behavior. The AW segmentation achieved the best results when using the prior tissue model (a fully supervised model). In all of the cases increasing the influence of the prior model improved the segmentation as compared to the unsupervised model.

The segmentation results show that the amount of falsely detected AW regions is considerably reduced when using the supervised models. This result is important in the context of automatic cervigrams analysis, where a small amount of false positives in the AW detection process is desirable. Still, the overall AW detection quality, as measured by the Dice metric, is as yet not satisfactory. Although considerably improved following the illumination correction process, there is still an overlap between the AW and the SE color distributions and the AW detection quality should be further improved.

The main conclusion extracted out of the pixel-based segmentation results is that
pixel-based color features are not distinctive enough for the tissues within the cervix, thus additional features should be used. This conclusion lead to the development of the superpixels-based segmentation framework. Using the superpixels for segmentation accounts for local continuity of pixel features within the image plane and augments the existing feature space with boundaries information and with local distributions of region features.

IV. Cervix tissues segmentation via clustering of superpixels

This part of the thesis presents a new framework appropriate for segmentation of elongated, non-convex regions within the cervix. It includes clustering of superpixels using both region and edge information. A new graph-cut criterion suitable for the special shape and color characteristics of regions within the cervix is introduced. The various components of the framework, including the superpixel representation, the superpixel similarity measure, the linkage measure used in the agglomerative process and different agglomerative clustering scenarios, are validated quantitatively.

The main conclusions drawn for the different steps within this framework are:

• The superpixels generation process is robust and the generated superpixels possess a satisfactory overlap with the cervix tissues across the different images.

• Different local scaling schemes that boost the differences between the values of the intra- and inter-tissues boundaries, were compared. The $\hat{e}_{ij}$ measure used in the current framework obtained the best results and provided the largest differences between the two boundary types. The local scaling process was shown to improve the results, as compared to the original boundaries (with no scaling).

• The quality of the similarity matrix was evaluated via its influence on the agglomerative clustering process. Three types of matrices were compared. The main differences between these matrices included the superpixel representation in color feature space: a Gaussian versus a $d$-dimensional histogram and the addition of a local scaling process to the region similarity measure. The local scaling process
was shown to improve the clustering results. Similar results were obtained with the two representations. The final framework uses the Gaussian distribution for superpixels representation, due to its lower memory consumption.

- The distribution of the weights within the two groups of inter and intra boundaries extracted from the entire image set, was shown to occupy the entire range of available values, with no clear distinction between the two groups. These distributions emphasize the complexity of the segmentation problem at hand, as no single threshold can be used across the images in order to distinguish between the two groups.

- Segmentation results with the simple agglomerative framework (Ag) were shown to provide the best results with the NMCut criterion (presented in this work). This criterion provides non-convex and elongated segments that possess a more accurate overlap with the cervix tissues in the majority of cases. The MinMax cut was shown to provide segments that are balanced in size but more convex in shape. The WMCut was shown to provide unbalanced clusters. Overall segmentation quality, in all examined cases was shown to decrease with the number of segments.

- The constrained agglomerative clustering (CAg) was suggested as an alternative to the Ag process that can determine a final number of segments per image. The CAg clustering prevents erroneous merging of segments using a stopping criterion, which is based on the original boundary weights. The results of the CAg clustering were shown to be highly dependent on the stopping criterion. The overall segmentation results of this algorithm were inferior to these of the other agglomerative algorithms. This might be related to the wide distribution of boundary weights, that discourage the use of a single threshold over the images. Future research should examine the option of a per-image threshold tuning, though this threshold might vary across the image itself.
• The agglomerative clustering with updates (AgU) was shown to be dependent on the predefined amount of components, \( N \), achieved prior to the updating step. The results were shown to improve with the increase of this number. This behavior reinforces the benefits stated for the algorithm. A comparison across the different cut criterions utilizing the AgU algorithm reviled a similar behavior as in the Ag case. The NMCut criterion obtained the best results.

• In a comparison between the different agglomerative algorithms: Ag, CAg, AgU, with the NMCut criterion and spectral clustering, the AgU obtained the best results (slightly better than these of the Ag algorithm). The segments generated by spectral clustering were shown to be balanced in size and convex in shape, similar to the Ag results with the MinMax-cut criterion.

• Segmentation results were evaluated on two test sets. AW lesions in \( \text{Set}_4 \) are clearly visible and a high level of agreement in their expert markings exist. Results on \( \text{Set}_4 \) were better throughout the different experiments, as compared to the results on \( \text{Set}_1 \) (with a random choice of cervigrams). This observation emphasizes the complexity of the segmentation task at hand.

In conclusion, the current work provides the following main contributions to the task of fully automated cervix tissues segmentation: 1) It explores the complexity of this task and the large variability of cases; 2) It presents local scaling schemes that cope with the large variability of features within a single tissue and the weak differences across tissues; 3) The work compares between different clustering algorithms, analyzes their behavior in the cervigrams scenario and points out their strengths and weaknesses. 4) Finally a new graph-cut criterion, which provides elongated, non-convex segments within the cervix; is presented and is shown to improve segmentation results in the majority of cases.

Although the current framework is promising, the overlap between the generated segments and the different tissues obtained acceptable results only when the number of segments was high (more than 50). Two future direction are considered: 1) In-
vestigating additional region-based features, to improve the distinction between the
different tissues, in order to obtain better segmentation results with a lower amount of
segments; 2) Examining the classification quality that can be obtained with a relatively
large amount of segments provided by the current framework. Preliminary experiments
in that direction are discussed next.

V. Preliminary classification results

In a preliminary classification experiment we compared segmentation results obtained
via different classification schemes and different image representation levels. Classifi-
cation was performed at the pixel, superpixel and segments level (obtained via clus-
tering of superpixels). Two classification schemes were examined: One is a based on
a Gaussian-KL (GKL) classifier, where tissues are modeled as single Gaussians and
the other is based on a K-nearest neighbor (Knn) classifier, where tissues are modeled
as a collection of Gaussians (Knn-GKL). The radial distance of each pixel from the
center of the cervix was used as an additional position feature in the classification pro-
cess. A comparison between the two frameworks for cervix tissues segmentation: The
pixel-based segmentation (Section 3.1) and segmentation via clustering of superpixels
(Section 3.2) was included. This comparison provides a summary for the current status
of the cervix tissues segmentation task.

The radial distance was shown to be an important feature for classification that
considerably improves the segmentation results. This result was more significant for
the CE tissue, which is known to be located around the cervix center and for the
SE tissue that surrounds it. Best AW classification results were obtained with the
Knn-GKL classifier with $K = 10$ neighbors, in the superpixel level. This result can
be explained as follows: The superpixel level benefits from the more robust features
of local distributions, thus outperforming the single pixels results. In addition, the
superpixels level doesn’t suffer from the erroneous merging generated during clustering,
which makes it better than the segments level. The Knn-GKL classifier provides better
results than the GKL classifier in that case because of the large overlap that exists
between the AW and SE Gaussian tissue models. This overlap is increased when the position feature is added. Example-based classification, as performed with the Knn-GKL classifier, might be more adequate in that case.

Best CE classification results were obtained in the segments level utilizing the model based GKL classifier. This can be explained by the more global integration of the textured CE regions, which is enforced by the segments level. Model based classification is adequate in that case, as this tissue is more concentrated in the center of the cervix and the position feature provides a good distinction between the CE and the other tissues. The best Accuracy and $SE_{Dice}$ results were obtained for the Knn-GKL classifier with the segments representation. The $AW_{Dice}$ and $CE_{Dice}$ results in this case were not significantly different from their best performance. Thus making the Knn-GKL classifier in the segments level adequate for the task of cervix tissues segmentation.

The last experiment evaluated the influence of the context-based rules of the pixel-based segmentation framework on the different classification results. This experiment included a comparison between the pixel-based segmentation framework and segmentation via clustering of superpixels. The best results overall were obtained by the segments classification via the Knn-GKL classifier. The context-based rules were found unnecessary in that case. A significant improvement was shown for the pixel-based segmentation when the context-based rules were added. This improvement was maintained also when the $R$ feature was added to the $Lb$ feature space. The overall classification results obtained in the different experiments should be further improved. Future work should include the optimization of additional classifiers, using more features (e.g. superpixel boundary information) and a larger training set.

### 6.2 General conclusions and future work

The main objectives of this research are to obtain an accurate segmentation of tissues within the cervix and identify important landmarks, from which different features can be extracted and used for future content-based indexing and retrieval of cervigrams.
A substantial effort was placed on the different tasks within the automated cervix analysis framework. Within these tasks, the detection of the cervix boundary obtained satisfactory results (mean $\text{Dice} = 0.81$, for $\text{Set}_1$). The illumination correction step was shown to reduce the amount of falsely detected AW regions and to improve the tissue segmentation quality.

The segmentation of cervix tissues, specifically the clinically important AW tissue, was shown to be a complex task for automatic analysis methods. A large overlap in feature space was shown to exist between the AW tissue and the other cervix tissues, a large variability of tissue boundaries was detected. The notion of AW segmentation complexity was reinforced by quantitatively measuring the variability of multiple experts markings, where the AW segmentation task was shown to be more complex than cervix boundary segmentation (Chapter 5). From the segmentation results obtained for merging of superpixels with $\text{Set}_1$ and $\text{Set}_4$ we can state the following: For AW lesions with strong agreement in expert markings ($\text{Set}_4$), segmentation results are acceptable (median $\text{AW}_{\text{Dice}} = 0.89$ for Ag clustering with 50 segments). For AW lesions with poor agreement in expert markings (as present in $\text{Set}_1$), results need to be improved further (median $\text{AW}_{\text{Dice}} = 0.57$, $\text{CE}_{\text{Dice}} = 0.73$). Preliminary tissue classification results of the generated segments were significantly improved with the addition of a radial position feature to the classification process. Applying a set of context-based rules was shown to further improve the results. Best classification results obtained in this work for $\text{Set}_1$ are: median $\text{SE}_{\text{Dice}} = 0.9$, $\text{CE}_{\text{Dice}} = 0.65$ and $\text{AW}_{\text{Dice}} = 0.37$.

The above observations lead to the following main options for future work:

- Use the generated segments of the current research for classification. Improve the classification schemes of the current work utilizing additional features, a larger training set and other classifiers.

- Explore ways in which tissues layout information can be more accurately incorporated into the segmentation or classification frameworks. This task requires the generation of an anatomical atlas for cervigrams, such as the probabilistic
atlases used for brain MRI (e.g. Warfield et al. [93]). These atlases were found to be indispensable for brain MRI segmentation, where a large overlap in the tissues intensities is known to exist. This task is very complex in the cervigrams scenario as appearance of a healthy cervix varies over time and the women parity and hormonal status. The variability among AW lesions size, position and color is very large. A concept that uses several models for the healthy cervix appearance should be considered in that case.

- Explore additional features that might provide more distinctive tissue models to guide the unsupervised segmentation process.

- Add a simple user intervention in order to aid the automatic analysis process. An example of such intervention might be the selection of a single point from each of the important tissues in the image and the inclusion of this information in the automated segmentation process. With the addition of such intervention, we might be neglecting our main goal of automated content-based access to the NIH database, but in terms of a clinical research, where accurate segmentations are required, this might be helpful.


Appendix A

Selection of pixel features for cervix tissues representation

Selecting an appropriate feature space is vital in order to represent and segment the cervix tissues. In this appendix we present a comparison between several feature spaces using the pixel-based segmentation framework and supervised modeling (Section 3.1). The results of this experiment lead to the selection of the \( Lb \) feature space, which is then used throughout the thesis.

We use a set of 35 manually marked cervigrams that were randomly selected out of \( Set_1 \) (Table 4.1) to learn a prior Gaussian model for each of the tissues in each of the examined feature spaces. A similar amount of samples is extracted per tissue out of the different tissue markings across the images and is used as a training set for the tissue modeling. Probabilistic segmentation (Equation (3.1)) using the prior models is performed on a test set that includes the remaining 83 images within \( Set_1 \). Results are evaluated within the cervix region as marked by the medical expert, following the preprocessing steps of SR detection and illumination correction.

The feature spaces used are different channel combinations from the \( CIE - Lab \), \( HSV \) and \( RGB \) color spaces (Gonzalez and Woods [23]). Two additional combinations were tested: one combination is \( G/S \) (the green channel from the \( RGB \) color space divided by saturation from the \( HSV \) color space). This combination is used in the
literature for AW segmentation (Lange [51]). The second combination is the \( LbT \) combination, where a multi-scale texture-contrast feature, \( T \), is added (Carson et al. [4]). This texture feature was used in our earlier works for segmentation of the CE tissue (Gordon et al. [28]). The texture feature was added only to the \( Lb \) combination following preliminary experiments with color features alone, where this combination attained the best results.

Following is a short description of the multi-scale texture-contrast feature. This feature describes both the underlying texture parameters and the adequate texture scale. It accounts for textured and non-textured regions and doesn’t consider a specific texture pattern, thus making it adequate for the detection of textured regions within the cervix. The scale is defined as the width of the Gaussian window within which gradient vectors of the image are pooled. The second moment matrix for the vectors within this window, computed about each pixel in the image, can be approximated using:

\[
M_\sigma(x, y) = G_\sigma(x, y) \ast (\nabla I)(\nabla I)^T, \tag{A.1}
\]

where \( G_\sigma \) is a separable binomial approximation to a Gaussian smoothing kernel with variance \( \sigma^2 \), and \( (\nabla I) \) is the gradient of the image intensity. Two texture descriptors are extracted for each pixel: polarity and texture-contrast. Polarity is a measure of the extent to which the gradient vectors in a certain neighborhood all point in the same direction, defined as:

\[
p_\sigma = \frac{|E_+ - E_-|}{E_+ + E_-}, \tag{A.2}
\]

where \( \sigma \) is the scale. The definitions of \( E_+ \) and \( E_- \) are:

\[
E_+ = \sum_{x,y} G_\sigma(x, y)[\nabla I \cdot \hat{n}]_+, \tag{A.3}
\]

\[
E_- = \sum_{x,y} G_\sigma(x, y)[\nabla I \cdot \hat{n}]_-, \tag{A.3}
\]

where \([\cdot]_+\) and \([\cdot]_-\) are the rectified positive and negative parts of their argument and
\( \hat{n} \) is a unit vector perpendicular to \( \phi \) (the direction of principal eigenvector of the second moment matrix, as defined in Equation (A.1)). This feature is used later for the selection of an appropriate texture scale for each pixel in the image. The contrast relates to the energy of the gradients in the vicinity of each pixel as given by Equation (A.4), where \( \lambda_1 \) and \( \lambda_2 \) are eigenvalues of \( M_\sigma \) (\( \lambda_1 \geq \lambda_2 \)):

\[
contrast = 2\sqrt{\lambda_1 + \lambda_2}.
\]  

(A.4)

The process of selecting an appropriate scale is based on the derivative of the polarity with respect to the scale. For each pixel \((x, y)\) the scale is selected as the first value for which the difference between values of polarity at successive scales is less then 2%. At the end of the texture features extraction phase, each pixel is associated with a contrast feature of the appropriate scale. The process is performed on the preprocessed cervigrams with filled-in SR regions.

Figure A.1 presents box plots for the Accuracy results computed over the entire test set, for each of the examined feature spaces. The \( Lb \) feature space (marked by a rectangle) attains the best results in this test. It possesses a significantly higher median value than the other feature spaces, with a relatively narrow interquartile range. Adding the texture feature in the \( LbT \) feature space doesn’t improve the results. The \( G/S \) combination achieved inferior results.

From a physiological point of view, it is not surprising that the \( Lb \) features are well suited to describe the cervix tissues, as it is adequate for describing the AW lesions. The large nucleus, characterizing the AW cells, with the additional cell dehydration caused by applying the acetic acid to the cervix surface, produce an increase in reflectance over all optical wave-lengths (Pogue et al. [73], Hoskings et al. [34]). This causes the whitening appearance of the AW tissue, which is best captured by the intensity feature, \( L \).

In order to explain the contribution of the \( b \) channel, let us look at the major chromophore in soft tissues: the blood hemoglobin. Blood hemoglobin is known to
diminish the blue-green absorption [73]. Variations in epithelial cell layer thickness or light scattering, appear as changes in color relative to the normal tissue. The abnormal AW cells make it difficult for light to pass through them, thus the blue-green absorption of the blood (which is underlying the epithelial cell layers) is diminished. The important role of the blue-green features in cervix cancer has been described by Pogue et al. [74]. Cannel $b$ represents the combinations of green and blue colors (Gonzalez and Woods [23]), and as such it can contribute to AW segmentation.
Appendix B

Selection of a superpixel region-based representation

One of the main factors that influence the clustering quality of a given graph, is the way graph weights are computed. The weights quality is influenced by the superpixels region and edge features distributions and the similarity measure used. The current experiment looks for a superpixels representation that provides the best distinction between the main cervix tissues. This is done based on region-features distributions, where the focus is placed on different features combinations. The best representation is selected via a superpixel classification test, in which superpixels are classified into one of three models representing the tissues of interest: AW, CE and SE.

The prior tissue models are learned from the same training set used in Appendix A. Results are analyzed within the automatically detected cervix boundary (using the data-driven curve evolution, Section 2.2.1) over the images of Set1 (Table 4.1). The experiment includes a comparison between two main classification schemes that use different feature representations. In scheme-1 superpixels are represented by histograms and in scheme-2 they are represented as Gaussians in a $d$-dimensional feature space. Following is a short description of each scheme:

1. In scheme-1 each superpixel is represented as a $d$-dimensional histogram, with 25 bins per dimension. Tissue models are represented in a similar way. The color
feature spaces used are the \textit{Lab} color space and the following \textit{HSV}-based color space: \(VS_1S_2\), where \(S_1 = SV\cos(2\pi H), S_2 = SV\sin(2\pi H)\) (Belongie \textit{et al.} [2]). Two additional combinations are tested: the \(Lb\) combination from the \textit{CIE- Lab} color space and the \(LbT\) combination, where the texture contrast feature (Carson \textit{et al.} [4]) is added, similar to Appendix A. The different features are linearly normalized in all of the cases. Two histogram-based similarity measures are used for classification: the Histogram Intersection (HI) and the discrete KL divergence. Given a superpixel histogram representation, \((p_1, ..., p_n)\), and a tissue model histogram representation, \((q_1, ..., q_n)\), the HI similarity measure is defined as: \[\sum_{k=1}^{n} \min(p_k, q_k)\] and the discrete KL divergence is defined as: \[\sum_{k=1}^{n} p_k \log \frac{p_k}{q_k}\].

2. In scheme-2 each superpixel is represented with a Gaussian in a \(d\)-dimensional feature space. Superpixels are classified into one of three Gaussian tissue models using the KL similarity measure between two Gaussians (Equation (3.6)). The feature spaces used include the \textit{Lab} and \(VS_1S_2\) color spaces and their three possible two-folds combinations. Texture was added to the \(Lb\) color combination as in the previous scheme.

All the experiments use tissue segmentation to evaluate the classification results (as in Section 4.5). Segmentation quality of the three tissues is evaluated next by a comparison to the expert markings using the Accuracy overlap measure. Figure B.1 presents box-plot Accuracy results computed over \textit{Set}_1 with the different classification schemes. Box-plots associated with scheme-1 are labeled 1-8. The discrete KL divergence is used in box-plots: 1-4 and the HI similarity measure in box-plots: 5-8. The box-plots (from left to right) within each of these sets, present results for the \(VS_1S_2\), \textit{Lab}, \(Lb\), \(LbT\) feature spaces, respectively. The different box-plots associated with scheme-2 are labeled: 9-17. The box-plots in this set are ordered from left to right according to the following feature combinations: \textit{Lab}, \(Lb\), \(La\), \(ab\), \(VS_1S_2\), \(VS_1\), \(VS_2\), \(S_1S_2\), \(LbT\). The cases with best segmentation Accuracy are: 10 (\(Lb\)) and 17 (\(LbT\)) that use scheme-2 for classification (marked by a rectangle).
Figure B.1: A comparison of superpixels representation quality via classification. Overall segmentation Accuracy results ($Set_1$). Best results are marked by a rectangle.

Figure B.2: A comparison of superpixels representation quality via classification. Amount of miss detections for AW (left) and CE(right) tissues ($Set_1$).
An additional measure that can be used to compare between the different classification schemes is the amount of miss-detections per tissue. Miss-detections are defined as cervigrams where the Similarity of the corresponding tissue is close to zero. Figure B.2 shows the amount of cervigrams where AW (left) and CE (right) tissues were missed. Cases 10 and 17, that obtained the best segmentation quality results in the previous analysis, obtain a relatively low amount of miss detections for both tissues. Following these experiment results, the $d$-dimensional Gaussian distribution in $Lb$ feature space is used for the superpixels region-based representation. The addition of the texture feature is not shown to provide a significant improvement.